Docket No.: PC-0044 CIP

$AF/646/1$

Response Under 37 C.F.R. 1.116 – Expedited Procedure
Examining Group 1646

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

Corres. and Mail

BOX AF

In re Application of: Bandman et al.

Title: HUMAN GPCR PROTEINS

Serial No.:   09/895,686   Filing Date: June 28, 2001

Examiner: O'Hara, E.   Group Art Unit: 1646

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

TRANSMITTAL FEE SHEET

Sir:

Transmitted herewith are the following for the above-identified application:

1. Return Receipt Postcard;
2. Brief on Appeal, including Appendix (42) pp., in triplicate);
3. Six (6) References, as cited in Brief (1 – 6);
4. Declaration of Tod Bedilion, Ph.D., Under 37 C.F.R. §1.132 (3 pp., in triplicate);
5. Declaration of Vishwanath R. Iyer, Ph.D., Under 37 C.F.R. §1.132 (5 pp., in triplicate);
6. Exhibits A – E, as cited in Iyer Declaration; and (in triplicate)
7. Ten (10) references published before the filing date of the instant application (References 9A – 9J). (in triplicate)

The fee has been calculated as shown below:.

☐	No additional Fee is required.	$	
☒	Fee for filing a Brief in support of an Appeal under 37 C.F.R. §1.17(c):	$	330.00
☐	Fee for Petition for Extension of Time Under 37 C.F.R. § 1.17(a):	$	.00
☒	Please charge Deposit Account No. **09-0108** in the amount of:	$	**330.00**

The Commissioner is hereby authorized to charge any additional fees required under 37 C.F.R. § 1.16 and

§ 1.17, or credit overpayment to Deposit Account No. **09-0108**. **A duplicate copy of this sheet is enclosed.**

Respectfully submitted,
INCYTE CORPORATION

Date: January 7, 2004

David G. Streeter
Reg. No. 43,168
Direct Dial Telephone: (650) 845-5741

Customer N .: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555 or Fax: (650) 849-8886

Doc No.118162   1   09/895,686

Docket No.: PC-0044 CIP

Response Under 37 C.F.R. 1.116 - Expedited Procedure
Examining Group 1646

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re Application of: Bandman et al.

Title: HUMAN GPCR PROTEINS

Serial No.:	09/895,686	Filing Date:	June 28, 2001
Examiner:	O'Hara, E	Group Art Unit:	1646

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

BRIEF ON APPEAL

Sir:

Further to the Notice of Appeal filed November 5, 2003, and received by the USPTO on November 7, 2003, herewith are three copies of Appellants' Brief on Appeal. Authorized fees include the $ **330.00** fee for the filing of this Brief.

This is an appeal from the decision of the Examiner finally rejecting claims 1-6 of the above-identified application.

01/09/2004 JADDO1   00000092 090108   09895686
01 FC:1402   330.00 DA

(1) REAL PARTY IN INTEREST

The above-identified application is assigned of record to Incyte Pharmaceuticals, Inc., (now Incyte Corporation, formerly known as Incyte Genomics, Inc.) (Reel 012272, Frame 0191) which is the real party in interest herein.

(2) RELATED APPEALS AND INTERFERENCES

Appellants, their legal representative and the assignee are not aware of any related appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the instant appeal.

(3) STATUS OF THE CLAIMS

Claims rejected: Claims 1-6.

Claims allowed: (none).

Claims canceled: Claims 13-20.

Claims withdrawn: Claims 7-12.

Claims on Appeal: Claims 1-6 (A copy of the claims on appeal, as amended, can be found in the attached Appendix.)

(4) STATUS OF AMENDMENTS AFTER FINAL

There were no amendments submitted after Final Rejection.

(5) SUMMARY OF THE INVENTION

Appellants' invention is directed to polynucleotides encoding a human G-protein coupled receptor (GPCR), SEQ ID NO:1, in particular, a metabotropic glutamate GPCR, based on the conservation of various sequence motifs characteristic of this family of proteins, in particular, the seven hydrophobic transmembrane domains characteristic of GPCRs. See specification, at page 11 and Table 1 and Figure 1. The glutamate GPCRs are described in the specification and the art of record as important in neurotransmission and involved in neurological disorders such as epilepsy, stroke, and neurodegeneration. See specification, at page 2. Polynucleotides encoding SEQ ID NO:1 are also disclosed as differentially expressed in thyroid tumors, in particular, follicular carcinoma based on Northern analysis in thyroid tissues. See specification, at page 35. The claimed polynucleotides are asserted to be useful in the diagnosis, treatment, and evaluation of therapies for neurological and neoplastic disorders, in particular, follicular carcinoma.

(6) <u>ISSUES</u>

1.   Whether claims 1-6 directed to SEQ ID NO:1 encoding polynucleotides meet the utility requirement of 35 U.S.C. §101. In particular, whether the conservation of sequence motifs and domains between the protein coded for by the claimed polynucleotide and metabotropic GPCRs, known to have utility in neurotransmission and neurological disorders, demonstrates a "substantial likelihood" of utility under 35 U.S.C. § 101. Whether there is evidence that the differential expression of the polynucleotide encoding SEQ ID NO:1 in thyroid tumors provides a substantial likelihood of utility for the claimed polynucleotides in the detection and diagnosis of thyroid tumors.

2.   Whether one of ordinary skill in the art would know how to use the claimed polynucleotides, e.g., in toxicology testing, drug development, and the diagnosis of disease, so as to satisfy the enablement requirement of 35 U.S.C. §112, first paragraph.

3.   Whether fragments and variants of the polynucleotides encoding SEQ ID NO:1 are sufficiently described in the specification that the skilled artisan would recognize applicant's possession of them at the time the application was filed in accordance with 35 U.S.C. § 112, First Paragraph.

4.   Whether the claimed polynucleotides are sufficiently described in priority application Serial No. 09/516,513, filed September 17, 1998 to meet the requirements of 35 U.S.C. § 112, first paragraph and claim an effective priority date of September 17, 1998 with respect to the now claimed invention.

(7) <u>GROUPING OF THE CLAIMS</u>

As to Issue 1

All of the claims on appeal stand or fall together.

As to Issue 2

All of the claims on appeal stand or fall together

As to Issue 3

All of the claims on appeal stand or fall together

As to Issue 4

Claims 1 and 3-6 stand or fall together.

(8) APPELLANTS' ARGUMENTS

The rejection of claims 1-6 under 35 U.S.C. §§ 101 and 112, first paragraph is improper, as the inventions of those claims have a patentable utility as set forth in the instant specification, and/or a utility well known to one of ordinary skill in the art.

Claims 1-6 stand rejected under 35 U.S.C. §§ 101 and 112, first paragraph, based on the allegation that the claimed invention lacks patentable utility. The rejection alleges in particular that:

- the claimed invention is not supported by either a substantial and specific asserted utility or a well- established utility. None of the described uses are considered to be specific or substantial utilities for either the protein or encoding nucleic acid molecules. Methods such as identification of ligands, use to screen for homologous genes, use to identify chromosomes or chromosomal locations, use to recombinantly produce protein or to generate antibodies are considered general methods applicable to any protein and/or nucleic acid.

- Applicants assertion that the claimed polynucleotide can be used in cancer diagnosis, in particular follicular carcinoma of the thyroid, is unconvincing because the correlation between the expression of the polynucleotide and follicular carcinoma is based on one single library. The determination of a cancer marker must be based on studying results from considerable number of patients, and statistical analysis. See Guidelines for Marker Development by the National Cancer Institute (NCI).

The invention at issue is a polynucleotide corresponding to a gene that is expressed in humans. The novel polynucleotide codes for a polypeptide demonstrated in the patent specification to be a member of the class of glutmate GPCRs, whose biological functions include control of neuerotransmission. The claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which requires

knowledge of how the polypeptide coded for by the polynucleotide actually functions. The claimed invention can be used, for example, as a marker for cancers of the thyroid, in particular, follicular carcinoma. See specification, at page 35. As a result of the benefits of these uses, the claimed invention already enjoys significant commercial success.

Applicants have previously submitted a Declaration by Dr. John C. Rockett showing the many reasons why the use of the claimed polynucleotides in gene expression profiling studies in toxicology testing would be readily apparent to the skilled artisan at the time the application was filed.

Applicants further submit two additional expert Declarations by Dr. Vishwanath R. Iyer and Dr. Tod Bedilion under 37 C.F.R. § 1.132, with respective attachments, and ten (10) scientific references filed before the September 17, 1998 priority date of the instant application. The Rockett Declaration, Iyer Declaration, Bedilion Declaration, and the ten (10) references fully establish that, prior to the September 17, 1998 filing date of the parent Bandman '513 application, it was well-established in the art that:

> polynucleotides derived from nucleic acids expressed in one or more tissues and/or cell types can be used as hybridization probes -- that is, as tools -- to survey for and to measure the presence, the absence, and the amount of expression of their cognate gene;

> with sufficient length, at sufficient hybridization stringency, and with sufficient wash stringency -- conditions that can be routinely established -- expressed polynucleotides, used as probes, generate a signal that is specific to the cognate gene, that is, produce a gene-specific expression signal;

> expression analysis is useful, *inter alia*, in drug discovery and lead optimization efforts, in toxicology, particularly toxicology studies conducted early in drug development efforts, and in phenotypic characterization and categorization of cell types, including neoplastic cell types;

> each additional gene-specific probe used as a tool in expression analysis provides an additional gene-specific signal that could not otherwise have been detected, giving a more comprehensive, robust, higher resolution, statistically more significant, and thus more useful expression pattern in such analyses than would otherwise have been possible;

> biologists, such as toxicologists, recognize the increased utility of more comprehensive, robust, higher resolution, statistically more significant

results, and thus want each newly identified expressed gene to be included in such an analysis;

nucleic acid microarrays increase the parallelism of expression measurements, providing expression data analogous to that provided by older, lower throughput techniques, but at substantially increased throughput;

accordingly, when expression profiling is performed using microarrays, each additional gene-specific probe that is included as a signaling component on this analytical device increases the detection range, and thus versatility, of this research tool;

biologists, such as toxicologists, recognize the increased utility of such improved tools, and thus want a gene-specific probe to each newly identified expressed gene to be included in such an analytical device;

the industrial suppliers of microarrays recognize the increased utility of such improved tools to their customers, and thus strive to improve salability of their microarrays by adding each newly identified expressed gene to the microarrays they sell;

it is not necessary that the biological function of a gene be known for measurement of its expression to be useful in drug discovery and lead optimization analyses, toxicology, or molecular phenotyping experiments;

failure of a probe to detect changes in expression of its cognate gene does not diminish the usefulness of the probe as a research tool; and

failure of a probe completely to detect its cognate transcript in any single expression analysis experiment does not deprive the probe of usefulness to the community of users who would use it as a research tool.]

The Patent Examiner does not dispute that the claimed polynucleotide can be used as a probe in cDNA microarrays and used in gene expression monitoring applications. Instead, the Patent Examiner contends that the claimed polynucleotide cannot be useful without precise knowledge of its biological function, or the biological function of the polypeptide it encodes. But the law has never required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

In any event, as demonstrated by the Rockett Declaration, the Iyer Declaration, and the Bedilion Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polynucleotide in the absence of any knowledge as to the precise function of the

protein encoded by it. The uses of the claimed polynucleotide in gene expression monitoring applications are in fact independent of its precise biological function.

I.   The applicable legal standard

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

> An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention "is incapable of serving any beneficial end").

Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. *See Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a "nebulous expression" such as "biological activity" or "biological properties" that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be "substantial." *Brenner*, 383 U.S. at 534. A "substantial" utility is a practical, "real-world" utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a "well-established" utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examining Procedure at § 706.03(a). Only if there is no "well-established" utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566, 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

II. Toxicology Testing and disease diagnosis are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are "well-established" uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application's specification. These uses are explained, in detail, in the Rockett Declaration, Iyer Declaration, and Second Bedilion Declaration accompanying this brief or previously submitted. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

A.   **The use of the claimed SEQ ID NO:1 encoding polynucleotides for toxicology testing, drug discovery, and disease diagnosis are practical uses that confer "specific benefits" to the public**

The claimed invention has specific, substantial, real-world utility by virtue of its use in toxicology testing, drug development and disease diagnosis through gene expression profiling. These uses are explained in detail in the accompanying Rockett Declaration, Iyer Declaration, and Bedilion Declaration, the substance of which is not rebutted by the Patent Examiner. There is no dispute that the claimed invention is in fact a useful tool in cDNA microarrays used to perform gene expression analysis. That is sufficient to establish utility for the claimed polynucleotide.

In his Declaration, Dr. Rockett explains the many reasons why a person skilled in the art in 1998 would have understood that any expressed polynucleotide is useful for a number of gene expression monitoring applications, *e.g.*, in cDNA microarrays, in connection with the development of drugs and the monitoring of the activity of such drugs. (Rockett Declaration at, e.g., ¶¶ 10-18).

> It is my opinion, therefore, based on the state of the art in toxicology at least since the mid-1990s . . . that disclosure of the sequence of a new gene or protein, with or without knowledge of its biological function, would have been sufficient information for a toxicologist to use the gene and/or protein in expression profiling studies in toxicology.[1] [Rockett Declaration, ¶ 18.]

In his Declaration, Dr. Bedilion explains why a person of skill in the art in 1998 would have understood that any expressed polynucleotide is useful for gene expression monitoring applications using cDNA microarrays. (Bedilion Declaration, e.g., ¶¶ 4-7.) In his Declaration, Dr. Iyer explains why a person of skill in the art in 1998 would have understood that any expressed polynucleotide is useful for gene expression monitoring applications using cDNA microarrays, stating that "[t]o provide maximum versatility as a research tool, the microarray should include ☐ and as a biologist I would want my microarray to include ☐ each newly identified gene as a probe." (Iyer Declaration, ¶ 9.)

"Use of the words 'it is my opinion' to preface what someone of ordinary skill in the art would have known does not transform the factual statements contained in the declaration into opinion testimony." *In re Alton*, 37 USPQ2d 1578, 1583 (Fed. Cir. 1996).

In addition, Dr. Rockett explains in his Declaration that "there are a number of other differential expression analysis technologies that precede the development of microarrays, some by decades, and that have been applied to drug metabolism and toxicology research, including: (1) differential screening; (2) subtractive hybridization, including variants such as chemical cross-linking subtraction, suppression-PCR subtractive hybridization and representational difference analysis; (3) differential display; (4) restriction endonuclease facilitated analyses, including serial analysis of gene expression (SAGE) and gene expression fingerprinting and (5) EST analysis." (Rockett Declaration, ¶ 7.)

Nowhere does the Patent Examiner address the fact that, as described on pages 31-32 of the Bandman '513 application, the claimed polynucleotides can be used as highly specific probes in, for example, cDNA microarrays - probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotide is known to be expressed, its utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon* v. *Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 ("Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., they are useful in analyzing compounds)" (emphasis added)).

Literature reviews published shortly before the filing of the Bandman '513 application describing the state of the art further confirm the claimed invention's utility. Rockett et al. confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.

* * *

Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

* * *

Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal . . . . However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis in original)

Rockett et al., Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential, Xenobiotica 29:655-691 (July 1999) (Rockett Declaration, Exhibit C).

In another pre-September 1998 article, Lashkari et al. state explicitly that sequences that are merely "predicted" to be expressed (predicted Open Reading Frames, or ORFs)-- the claimed invention in fact is known to be expressed-- have numerous uses:

Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons-- they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay. (emphasis added)

Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947 (Aug. 1997) (Reference No. 1).

B.   The use of polynucleotides coding for polypeptides expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now "well-established"

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, e.g., as described by Bedilion, Rockett, and Iyer in their Declarations.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett et al., *supra*:

> Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Rockett Declaration, Exhibit C, page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999) (Reference No. 2); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000) (Reference No. 3).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

> for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip.

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. "Arrays are at their most powerful when they contain the entire genome of the species they are being used to study." John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, Environ. Health Perspec.107:681-685 (1999) (Reference No. 4). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Reference No. 5), indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

Further evidence of the well-established utility of all expressed polypeptides and polynucleotides in toxicology testing is found in U.S. Pat. No. 5,569,588 (Reference No. 9e) and published PCT applications WO 95/21944 (Reference No. 9a), WO 95/20681 (Reference No. 9b), and WO 97/13877 (Reference No. 9g).

WO 95/21944 ("Differentially expressed genes in healthy and diseased subjects"), published August 17, 1995, describes the use of microarrays in expression profiling analyses, emphasizing that *patterns* of expression can be used to distinguish healthy tissues from diseased tissues and that *patterns* of expression can additionally be used in drug development and toxicology studies, without knowledge of the biological function of the encoded gene product. In particular, and with emphasis added:

> The present invention involves . . . methods for diagnosing diseases . . . characterized by the presence of [differentially expressed] . . . genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/ polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST. . . . Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of diseases based on differential expression of genes of unknown function . . . . [abstract]

> The method [of the present invention] involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA

polynucleotide sequences . . . and a defined set of oligonucleotide/polynucleotide[]
. . . immobilized on a support. Those defined [immobilized]
oligonucleotide/polynucleotide sequences are <u>representative of the total expressed</u>
<u>genetic component of the cells</u>, tissues, organs or organism as defined by the
collection of partial cDNA sequences (ESTs). [page 2]

The present invention meets the unfilled needs in the art by providing
methods for the . . . <u>use of gene fragments and genes, even those of unknown full</u>
<u>length sequence and unknown function, which are differentially expressed</u> in a
healthy animal and in an animal having a specific disease or infection by use of
ESTs derived from DNA libraries of healthy and/or diseased/infected animals.
[page 4]

Yet another aspect of the invention is that it provides . . . a means for . . .
monitoring the efficacy of disease treatment regimes <u>including</u> . . . <u>toxicological</u>
<u>effects thereof</u>." [page 4]

It has been appreciated that one or more differentially identified EST or
gene-specific oligonucleotide/polynucleotides <u>define a pattern</u> of differentially
expressed genes diagnostic of a predisease, disease or infective state. <u>A knowledge</u>
<u>of the specific biological function of the EST is not required</u> only that the EST[]
identifies a gene or genes whose altered expression is associated reproducibly with
the predisease, disease or infectious state. [page 4]

As used herein, the term 'disease' or 'disease state' refers to any condition
which deviates from a normal or standardized healthy state in an organism of the
same species in terms of differential expression of the organism's genes. . .
[whether] of genetic or environmental origin, for example, an inherited disorder
such as certain breast cancers. . . .[or] administration of a drug or exposure of the
animal to another agent, e.g., nutrition, which affects gene expression. [page 5]

As used herein, the term 'solid support' refers to any known substrate which
is useful for the immobilization of large numbers of oligonucleotide/polynucleotide
sequences by any available method . . . [and includes, inter alia,] nitrocellulose, . . .
glass, silica. . . . [page 6]

By 'EST' or 'Expressed Sequence Tag' is meant a partial DNA or cDNA
sequence of about 150 to 500, more preferably about 300, sequential nucleotides. . .
. [page 6]

One or more libraries made from a single tissue type typically provide at
least about 3000 different (i.e., unique) ESTs and <u>potentially the full complement of</u>
<u>all possible ESTs representing all cDNAs e.g., 50,000 to 100,000 in an animal such</u>
<u>as a human</u>. [page 7]

The lengths of the defined oligonucleotide/ polynucleotides may be readily increased or decreased as desired or needed. . . . The length is generally guided by the principle that it should be of sufficient length to insure that it is on[] average only represented once in the population to be examined. [page 7]

Comparing the . . . hybridization patterns permits detection of those defined oligonucleotide/ polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions [of the solid support]. [page 13]

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained[;] rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used [to] form diagnostic patterns or to identify which particular EST is detected. For example, all known ESTs from an organism are used to produce a 'master' solid support to which control sample and disease samples are alternately hybridized. [page 14]

Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This[,] like many of the foregoing embodiments[,] may use known or unknown ESTs derived from many libraries. [page 18]

Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal[s], especially humans. [page 18]

WO 95/20681 ("Comparative Gene Transcript Analysis"), filed in 1994 by Appellants' assignee and published August 3, 1995, has three issued U.S. counterparts: U.S. Pat. Nos. 5,840,484, issued November 24, 1998; 6,114,114, issued September 5, 2000; and 6,303,297, issued October 16, 2001.

The specification describes the use of transcript expression *patterns,* or "images", each comprising multiple pixels of gene-specific information, for diagnosis, for cellular phenotyping, and in toxicology and drug development efforts. The specification describes a plurality of methods for obtaining the requisite expression data -- one of which is microarray hybridization -- and equates the uses of the expression data from these disparate platforms. In particular, and with emphasis added:

The invention provides a "method and system for quantifying the relative abundance of gene transcripts in a biological specimen. . . . [G]ene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is <u>correlated</u> to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides <u>a method for comparing the gene transcript image analysis</u> from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens." [abstract]

"<u>[W]e see each individual gene product as a 'pixel' of information</u> which <u>relates to the expression of that, and only that, gene</u>. We teach herein [] methods whereby <u>the individual 'pixels' of gene expression information can be combined into a single gene transcript 'image,'</u> in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood." [page 2]

"The present invention avoids the drawbacks of the prior art by providing <u>a method to quantify the relative abundance of multiple gene transcripts in a given biological specimen</u>. . . . The method of the instant invention provides for detailed diagnostic <u>comparisons of cell profiles</u> revealing numerous changes in the expression of individual transcripts." [page 6]

"High resolution analysis of gene expression be <u>used directly as a diagnostic profile</u>. . . ." [page 7]

"The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed." [page 7]

"The invention . . . includes a method of comparing specimens containing gene transcripts." [page 7]

"The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens." [i.e., the results yield analogous data to microarrays] [page 8]

"Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made." [page 8]

"In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities." [page 9]

"In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens. . . . " [page 9]

"[T]wo or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells." [pages 9 - 10]

"The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens. . . . This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as 'gene transcript image analysis' or 'gene transcript frequency analysis'. The present invention allows one to obtain a profile for gene transcription in any given population of cells or tissue from any type of organism." [page 11]

"The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few." [page 12]

"[G]ene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates." [page 12]

"For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues. . . ." [page 12]

"In toxicology, . . . [g]ene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. . . . " [page 12]

"In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate between cancer cells which respond to anti-cancer agents and those which do not respond." [page 12]

"In a further embodiment, comparative gene transcript frequency analysis is used . . . for the selection of better pharmacologic animal models." [page 14]

"In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to give a highly detailed gene transcript profile of a diseased state or condition." [page 14]

"<u>An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined.</u>" [page 15]

"<u>[T]his research tool</u> provides a way to get new drugs to the public faster and more economically." [page 36]

"<u>In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a clinical marker.</u>" [page 38]

"[T]he gene transcript changes noted in the earlier rat <u>toxicity study</u> are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript image analyses are evaluated as indicators of <u>toxicity by correlation</u> with clinical signs and symptoms and other laboratory results. . . . The . . . analysis highlights any toxicological changes in the treated patients." [page 39]

<u>U.S. Pat. No. 5,569,588</u> ("Methods for Drug Screening") ("the '588 patent"), issued October 29, 1996, with a priority date of August 1995, describes an expression profiling platform, the "genome reporter matrix", which is different from nucleic acid microarrays. Additionally describing use of nucleic acid microarrays, the 588 patent makes clear that the utility of comparing multidimensional expression datasets is independent of the methods by which such profiles are obtained. The 588 patent speaks clearly to the usefulness of such expression analyses in drug development and toxicology, particularly pointing out that a gene's failure to change in expression level is a useful result. Thus, with emphasis added,

The invention provides "[m]ethods and compositions for modeling the transcriptional responsiveness of an organism to a candidate drug. . . . [The final step of the method comprises] comparing reporter gene product signals for each cell before and after contacting the cell with the candidate drug to <u>obtain a drug response profile</u> which provides a model of the transcriptional responsiveness of said organism to the candidate drug." [abstract]

"The present invention exploits the recent advances in genome science to provide for the rapid screening of large numbers of compounds against a systemic target comprising <u>substantially all targets in a pathway [or] organism.</u>" [col. 1]

117940   18   09/895,686

"The ensemble of reporting cells comprises as comprehensive a collection of transcription regulatory genetic elements as is conveniently available for the targeted organism so as to most accurately model the systemic transcriptional response. Suitable ensembles generally comprise thousands of individually reporting elements; preferred ensembles are substantially comprehensive, i.e. provide a transcriptional response diversity comparable to that of the target organism. Generally, a substantially comprehensive ensemble requires transcription regulatory genetic elements from at least a majority of the organism's genes, and preferably includes those of all or nearly all of the genes. We term such a substantially comprehensive ensemble a genome reporter matrix." [col. 2]

"Drugs often have side effects that are in part due to the lack of target specificity. . . . [A] genome reporter matrix reveals the spectrum of other genes in the genome also affected by the compound. In considering two different compounds both of which induce the ERG10 reporter, if one compound affects the expression of 5 other reporters and a second compound affects the expression of 50 other reports, the first compound is, a priori, more likely to have fewer side effects." [cols. 2 - 3]

"Furthermore, it is not necessary to know the identity of any of the responding genes." [col. 3]

"[A]ny new compound that induces the same response profile as [a] . . . dominant tubulin mutant would provide a candidate for a taxol-like pharmaceutical." [col. 4]

"The genome reporter matrix offers a simple solution to recognizing new specificities in combinatorial libraries. Specifically, pools of new compounds are tested as mixtures across the matrix. If the pool has any new activity not present in the original lead compound, new genes are affected among the reporters." [col. 4]

" A sufficient number of different recombinant cells are included to provide an ensemble of transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug. In a preferred embodiment, the matrix is substantially comprehensive for the selected regulatory elements, e.g. essentially all of the gene promoters of the targeted organism are included." [cols. 6 – 7]

"In a preferred embodiment, the basal response profiles are determined. . . . The resultant electrical output signals are stored in a computer memory as genome reporter output signal matrix data structure associating each output signal with the coordinates of the corresponding microtiter plate well and the stimulus or drug. This information is indexed against the matrix to form reference response profiles that are used to determine the response of each reporter to any milieu in which a stimulus may be provided. After establishing a basal response profile for the matrix, each cell is contacted with a candidate drug. The term drug is used loosely

to refer to agents which can provoke a specific cellular response. . . . <u>The drug</u> <u>induces a complex response pattern</u> of repression, silence and induction across the matrix . . . .The response profile reflects the cell's transcriptional adjustments to maintain homeostasis in the presence of the drug. . . . After contacting the cells with the candidate drug, the reporter gene product signals from each of said cells is again measured to determine a stimulated response profile. The basal o[r] background response profile is then compared with . . . the stimulated response profile to identify the cellular response profile to the candidate drug." [cols. 7 - 8]

"<u>In another embodiment of the invention, a matrix [i.e., array] of</u> <u>hybridization probes corresponding to a predetermined population of genes of the</u> <u>selected organism is used</u> to specifically detect changes in gene transcription which result from exposing the selected organism or cells thereof to a candidate drug. In this embodiment, one or more cells derived from the organism is exposed to the candidate drug in vivo or ex vivo under conditions wherein the drug effects a change in gene transcription in the cell to maintain homeostasis. Thereafter, the gene transcripts, primarily mRNA, of the cell or cells is isolated . . . [and] then contacted with an ordered matrix [array] of hybridization probes, each probe being specific for a different one of the transcripts, <u>under conditions where each of the</u> <u>transcripts hybridizes with a corresponding one</u> of the probes to form hybridization pairs. The ordered matrix of probes provides, in aggregate, complements for an ensemble of genes of the organism sufficient to model the transcriptional responsiveness of the organism to a drug. . . . The <u>matrix-wide signal profile of the</u> <u>drug-stimulated cells is then compared with a matrix-wide signal profile of negative</u> <u>control cells to obtain a specific drug response profile.</u>" [col. 8]

"The invention also provides means for computer-based qualitative analysis of candidate drugs and unknown compounds. A wide variety of reference response profiles may be generated and used in such analyses." [col. 8]

"<u>Response profiles for an unknown stimulus (e.g. new chemicals, unknown</u> <u>compounds or unknown mixtures) may be analyzed by comparing the new stimulus</u> <u>response profiles with response profiles to known chemical stimuli.</u>" [col. 9]

"The response profile of a new chemical stimulus may also be compared to a known genetic response profile for target gene(s)." [col. 9]

The August 11, 1997 press release from the '588 patent's assignee, Acacia Biosciences

(now part of Merck) (reference "9h" attached hereto), and the September 15, 1997 news report by

Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," *Genetic Engineering*

News (reference "9i" attached hereto), attest the commercial value of the methods and technology

described and claimed in the '588 patent.

<u>**WO 97/13877**</u> ("Measurement of Gene Expression Profiles in Toxicity

Determinations"), published April 17, 1997, describes an expression profiling technology

differing somewhat from the use of cDNA microarrays and differing from the genome reporter

matrix of the '588 patent; but the use of the data is analogous. As per its title, the reference

describes use of expression profiling in toxicity determinations. In particular, and with emphasis

added:

> "[T]he invention relates to a method for detecting and monitoring changes
> in gene expression patterns in in vitro and in vivo systems for determining the
> toxicity of drug candidates." [Field of the invention]

> "An object of the invention is to provide <u>a new approach to toxicity</u>
> <u>assessment based on an examination of gene expression patterns, or profiles</u>, in in
> vitro or in vivo test systems." [page 3]

> "Another object of the invention is to provide a rapid and reliable method
> for correlating gene expression with short term and long term toxicity in test
> animals." [page 3]

> "The invention achieves these and other objects by providing a method for
> massively parallel signature sequencing of genes expressed in one or more selected
> tissues of an organism exposed to a test compound. An important feature of the
> invention is the application of novel . . . methodologies that permit the formation of
> gene expression profiles for selected tissues . . . . Such <u>profiles may be compared</u>
> with those from tissues of control organisms at single or multiple time points <u>to</u>
> <u>identify expression patterns predictive of toxicity</u>." [page 3]

> "As used herein, the terms 'gene expression profile,' and 'gene expression
> pattern' which is used equivalently, means a frequency distribution of sequences of
> portions of cDNA molecules sampled from a population of tag-cDNA conjugates. .
> . . Preferably, the total number of sequences determined is at least 1000; <u>more</u>
> <u>preferably, the total number of sequences determined in a gene expression profile is</u>
> <u>at least ten thousand</u>." [page 7]

> "The invention provides a method for determining the toxicity of a
> compound by analyzing changes in the gene expression profiles in selected tissues
> of test organisms exposed to the compound. . . . . Gene expression profiles derived
> from test organisms are compared to gene expression profiles derived from control
> organisms. . . . " [page 7]

Therefore, the potential benefit to the public, in terms of lives saved and reduced health

care costs, are enormous. Evidence of the benefits of this information include:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangiers disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.

- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.

- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be overturned regardless of their merit.

C.   **The uncontested fact that the claimed polynucleotide encodes a protein in the GPCR family also demonstrates utility**

In addition to having substantial, specific and credible utilities in numerous gene expression monitoring applications, it is undisputed that the claimed polynucleotide encodes for a protein having the sequence shown as SEQ ID NO:1 in the patent application. Appellants have demonstrated that SEQ ID NO:1 is a member of the GPCR family, and that the GPCR family of proteins includes glutamate GPCRs that function in neurotransmission, and play a role in certain neurological diorders.

The Patent Examiner does not dispute any of the facts set forth in the previous paragraph. Neither does the Patent Examiner dispute that, if a polynucleotide encodes for a protein that has a substantial, specific and credible utility, then it follows that the polynucleotide also has a

substantial, specific and credible utility.

The Examiner must accept the applicant's demonstration that the polypeptide encoded by the claimed invention is a member of the GPCR family and that utility is proven by a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

Nor has the Examiner provided any evidence that any member of the GPCR family, let alone a substantial number of those members, is not useful. In such circumstances, the only reasonable inference is that the polypeptide encoded by the claimed invention must be, like the other members of the GPCR family, useful.

D. Objective evidence corroborates the utilities of the claimed invention

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. "Real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing the sequences of all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequence. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Appellants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the claimed sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's

databases have proven to be valuable in, for example, the identification and development of drug candidates. Page et al., in discussing the identification and assignment of candidate drug targets, state that "rapid identification and assignment of candidate targets and markers represents a huge challenge ... [t]he process of annotation is similarly aided by the quantity and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals)" Page, M.J. et al., "Proteomics: a major new technology for the drug discovery process," Drug Discov. Today 4:55-62 (1999) (Reference No. 6), see page 58, col. 2). As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's invention of the claimed polynucleotide and its use of that polynucleotide on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

III. The Patent Examiner's rejections are without merit

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that the disclosed and well-established utilities for the claimed polynucleotide are not "specific, substantial, and credible" utilities. (Final Office Action at page 3). The Examiner is incorrect both as a matter of law and as a matter of fact.

A. The precise biological role or function of an expressed polynucleotide is not required to demonstrate utility

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological role" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a cDNA microarray to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on

117940   24   09/895,686

biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an "identifiable benefit" in presently available form. *Juicy Whip Inc.* v. *Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the First Bedilion Declaration (at, *e.g.*, ¶¶ 10 and 15), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called "throwaway" utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged as much (66 F.R. at 1095):

> [T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed nucleic acid, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

B.   Membership in a class of useful products can be proof of utility

Despite the uncontradicted evidence that the claimed polynucleotide encodes a polypeptide in the GPCR family, the Examiner refused to impute the utility of the members of the GPCR family to SEQ ID NO:1. In the Final Office Action, the Patent Examiner takes the

position that, unless Appellants can identify which particular biological function within the class of GPCRs is possessed by SEQ ID NO:1, utility cannot be imputed. See Final Office Action, page 4. To demonstrate utility by membership in the class of GPCRs, the Examiner would require that all GPCRs possess a "common" utility.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility, and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether or not the members of the class possess one utility or many. See *Brenner* v. *Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a "general" class is insufficient to demonstrate utility only if the class contains a sufficient number of useless members such that a person of ordinary skill in the art could not impute utility by a substantial likelihood. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g., Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).

The Examiner addresses GPCRs as if the general class in which it is included is not the GPCR family, but rather all polynucleotides or all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these "general classes" may contain a substantial number of useless members, the GPCR family does not. The GPCR family is sufficiently specific to rule out any reasonable possibility that SEQ ID NO:1 would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the GPCR class of signaling molecules has any, let alone a substantial number, of useless members, the Examiner must conclude that there is a "substantial likelihood" that the SEQ ID NO:1 encoded by the claimed polynucleotide is useful. It follows that the claimed polynucleotide also is useful.

      C.   **Because the uses of the claimed polynucleotide in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the**

invention itself, the claimed invention has substantial utility

The Examiner's rejection of the claims at issue as not having a "substantial" use is tantamount to a rejection based on an allegation that the only use of the claimed invention is as a tool for further research. Because the PTO's rejection assumes a substantial overstatement of the law, and is incorrect in fact, it must be overturned.

There is no authority for the proposition that use as a tool for research is not a substantial utility. Indeed, the Patent Office has recognized that just because an invention is used in a research setting does not mean that it lacks utility (Section § 2107.01 of the Manual of Patent Examining Procedure, 8th Edition, August 2001, under the heading I. Specific and Substantial Requirements, Research Tools):

> Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact "useful" in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified utility and inventions whose specific utility requires further research to identify or reasonably confirm.

The Patent Office's actual practice has been, at least until the present, consistent with that approach. It has routinely issued patents for inventions whose only use is to facilitate research, such as DNA ligases. These are acknowledged by the PTO's Training Materials themselves to be useful, as well as DNA sequences used, for example, as markers.

Only a limited subset of research uses are not "substantial" utilities: those in which the only known use for the claimed invention is to be an **object** of further study, thus merely inviting further research. This follows from *Brenner*, in which the U.S. Supreme Court held that a process for making a compound does not confer a substantial benefit where the <u>only</u> known use of the compound was to be the object of further research to determine its use. *Id.* at 535. Similarly, in *Kirk*, the Court held that a compound would not confer substantial benefit on the public merely because it might be used to synthesize some other, unknown compound that would confer substantial benefit. *Kirk*, 376 F.2d at 940, 945 ("What appellants are really saying to those in the art is take these steroids, experiment, and find what use they do have as medicines."). Nowhere do those cases state or imply, however, that a material cannot be patentable if it has some other beneficial use in research.

D. The Patent Examiner failed to demonstrate that a person of ordinary skill in the art would reasonably doubt the utility of the claimed invention

The Examiner alleges that applicants asserted use of the claimed polynucleotide in the detection and diagnosis of cancer, in particular, thyroid cancer, is based on a correlation with thyroid cancer in on a single library representing follicular carcinoma of the thyroid. See specification, at page 35. Applicants reiterate that the asserted utility for the polynucleotide encoding SEQ ID NO:1 in the detection and diagnosis of follicular carcinoma of the thyroid, based on a significant (4-fold) differential expression in that disease condition, is both specific, substantial, and credible. The Examiners' allegation that the asserted utility is not credible because it is based on expression of the transcript in only one library ignores the fact that a number of thyroid libraries were examined representing both normal and diseased thyroid, and that only libraries associated with thyroid cancer were found to express the gene. In particular, the gene was most highly expressed in a thyroid follicular carcinoma tumor library (THYRTUP02), but was also expressed in a library associated with follicular adenoma (THYRNOT03), a precancerous condition to follicular carcinoma. Such evidence provides more than a "substantial likelihood" that the polynucleotide may be used in the detection and diagnosis of the disease. Further, the evidence provided from the Northern analysis for SEQ ID NO:7 supports applicants assertion or the use of the claimed polynucleotide in cancer as disclosed in the Bandman '513 priority application at pages 29-30. The Examiners' reliance on references such as the NCI Guidelines for Marker Development to support her position is merely an attempt to raise the standard for utility to one of near certainty. However, the standard applicable in this case is not proof to certainty, but rather proof to reasonable probability. *Brenner*, 383 U.S. at 532.

Applicants' Showing of Facts Overcomes The Examiner's Concern That Applicants' Invention Lacks "Specific Utility"

The Examiner alleges that the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility. (Final Office Action, page 3.)

Appellants' submission of additional facts overcomes this concern. Those facts demonstrate that, far from applying *regardless* of the specific properties of the claimed

invention, the utility of Appellants' claimed polynucleotides as gene-specific probes *depends upon* specific properties of the polynucleotides, that is, their nucleic acid sequences.

"[E]ach probe on . . . [a "high density spotted microarray[]"], with careful design and sufficient length, and with sufficiently stringent hybridization and wash conditions, *binds specifically* and with minimal cross-hybridization, to the probe's cognate transcript" [1], "[e]ach gene included as a probe on a microarray provides *a signal that is specific to the cognate transcript*, at least to a first approximation." [2] Accordingly, "each additional probe makes an additional transcript newly detectable by the microarray, increasing the detection range, and thus versatility, of this analytical device for gene expression profiling" [3], equally, "[e]ach new gene-specific probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device." [4]

Although not required for present purposes, it would be appropriate to state on the record here that the specificity of nucleic acid hybridization was well-established far earlier than the development of high density spotted microarrays in 1995, and indeed is the well-established underpinning of many, perhaps most, molecular biological techniques developed over the past 30 - 40 years.

IV. By requiring the patent applicant to assert a particular or unique utility, the Patent Examination Utility Guidelines and Training Materials applied by the Patent Examiner misstate the law

There is an additional, independent reason to overturn the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website www.uspto.gov, March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

---

[1] Declaration of Dr. John C. Rockett, ¶ 10(i), emphasis added.

[2] Declaration of Dr. Vishwanath R. Iyer, ¶ 7 (emphasis added). See the footnote at ¶ 7 for a slightly more "nuanced" view.

[3] Declaration of Dr. John C. Rockett, ¶ 10(ii).

[4] Declaration of Dr. Vishwanath R. Iyer, ¶ 7.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: "specific" utilities which meet the statutory requirements, and "general" utilities which do not. The Training Materials define a "specific utility" as follows:

> A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as "gene probe" or "chromosome marker" would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between "specific" and "general" utilities by assessing whether the asserted utility is sufficiently "particular," *i.e.*, unique (Training Materials at page 52) as compared to the "broad class of invention." (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, Written Description Guidelines and Utility Guidelines, 82 J.P.T.O.S. 77, 97 (Feb. 2000) ("With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.")).

Such "unique" or "particular" utilities never have been required by the law. To meet the utility requirement, the invention need only be "practically useful," *Natta*, 480 F.2d 1 at 1397, and confer a "specific benefit" on the public. *Brenner*, 383 U.S. at 534. Thus, incredible "throw-away" utilities, such as trying to "patent a transgenic mouse by saying it makes great snake food," do not meet this standard. Karen Hall, Genomic Warfare, The American Lawyer 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where "specific utility" is defined (page 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be "definite," not particular. *Montedison,* 664 F.2d at 375. Appellants are not aware of any court that has rejected an assertion of utility on the grounds that it is not "particular" or "unique" to the specific invention. Where courts have found utility to be too "general," it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a

specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had "useful biological activity" was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a "particular" type of cancer was determined to satisfy the specificity requirement). "Particularity" is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. *Montedison*, 664 F.2d at 374-75.

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of "general" utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied, they would render unpatentable whole categories of inventions that heretofore have been considered to be patentable and that have indisputably benefitted the public, including the claimed invention. See *supra* § II.B. Thus the Training Materials cannot be applied consistently with the law.

V.   **To the extent the rejection of the claimed invention under 35 U.S.C. § 112, first paragraph, is based on the improper rejection for lack of utility under 35 U.S.C. § 101, it must be reversed.**

The rejection set forth in the Office Action is based on the assertions discussed above, i.e., that the claimed invention lacks patentable utility. To the extent that the rejection under 35 U.S.C. § 112, first paragraph, is based on the improper allegation of lack of patentable utility under 35 U.S.C. § 101, it fails for the same reasons.

CONCLUSION

Appellants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of "lack of specificity," as set forth in the Office Action and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Neither are they scientifically correct, nor supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, "like a nose of wax," [5] to target rejections of claims to polypeptide and polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be reversed.

Moreover, to the extent the above rejections were based on the Revised Interim and final Examination Guidelines and Training Materials, those portions of the Guidelines and Training Materials that form the basis for the rejections should be determined to be inconsistent with the law.

Claims 1-6 stand rejected under 35 U.S.C. § 112, first paragraph, as containing subject matter which is not described in the specification in such a was as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention. The rejection alleges in particular, that:

• while the specification describes a polypeptide sequence consisting of SEQ ID NO:1, the claims encompass polypeptides comprising fragments and homologues that vary

---

[5] "The concept of patentable subject matter under §101 is not 'like a nose of wax which may be turned and twisted in any direction * * *.' *White v. Dunbar*, 119 U.S. 47, 51." *(Parker v. Flook,* 198 USPQ 193 (US SupCt 1978))

substantially in length and also in amino acid composition. The instant disclosure of a single polypeptide, that of SEQ ID NO:1, does not support the scope of the claimed genus, which encompasses a substantial variety of subgenera. See *Reagents of the University of California v Eli Lilly* with respect to the premise that "A description of a genus of cDNAs may be achieved by means of a recitation of a representative number of cDNAs, defined by nucleotide sequence, falling within the scope of the genus, or a recitation of structural features common to the genus, which features constitute a substantial portion of the genus". The Examiner then cited various references alleging to support the unpredictability of protein function based on sequence homology. See, in particular, Vukicevic et al.; Tischer et al.; and Kopchick et al. The Examiner concluded by saying that given the unpredictability of homology comparisons, and the fact that the specification fails to provide objective evidence that the additional sequences are indeed species of the claimed genus, it cannot be established that a representative number of species have been disclosed by the claims. Further, the Examiner stated, no activity is set forth for the additional sequences.

The recited fragments and variants of SEQ ID NO:1 and SEQ ID NO:2 are sufficiently described in chemical and structural terms that the skilled artisan would recognize applicant's possession of them at the time the application was filed

With respect to fragments of SEQ ID NO:1, as recited in claim 1, applicants submit that the recited fragments are disclosed in the specification and claims in terms of their specific amino acid sequences and therefore clearly meet the requirements for written description under 35 U.S.C. § 112, first paragraph..

The claimed "homologues" of SEQ ID NO:1 referred to by the Examiner presumably relate to variants of SEQ ID NO:1 and SEQ ID NO:7, as recited in claims 1 and 2, respectively. Applicants submit that the polypeptides and polynucleotides of the invention, including the recited variants, are adequately described in accordance with 35 U.S.C. § 112, first paragraph, and supported by relevant case law, some of which is referred to by the Examiner.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law.

. . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention.* The invention is, for purposes of the "written description" inquiry, *whatever is now claimed. Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

Attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that :

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail. If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met.

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 and SEQ ID NO:7 are specifically disclosed in the priority application Serial No. 09/156,513 (see, for example, page 2, lines 34-37 and page 3, lines 13-14). Variants of SEQ ID NO:1 and SEQ ID NO:7 are described, for example, at page 2, line 38 through page 3, line 2. In particular, the preferred, more preferred, and most preferred variants (80%, 90%, and 95% amino acid sequence similarity to SEQ ID NO:1) are described, for example, at page 12, lines 13-16 of priority application Serial No. 09/156,513. Incyte clones in which the nucleic acids encoding the human HGPRP-1 (SEQ ID NO:1) were first identified and libraries from which those clones were isolated are described, for example, at page 11, lines 24-30 and Table 1 of the priority application. Chemical and structural features of SEQ ID NO:1 are described, for example, on page 11, lines 31-35 and Table 2 of the priority application. Given SEQ ID NO:1, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:1

having at least 90% sequence identity to SEQ ID NO:1. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

A. The Specificati n pr vides an adequate written description of the claimed "variants" f SEQ ID NO:1.

The Office Action has further asserted that the claims are not supported by an adequate written description because:

> Claims 1-6 contain "subject matter which is not described in the specification in such a was as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention".

(page 8 of the Final Office Action)

Such a position is believed to present a misapplication of the law.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which "DNA claims" have been at issue (which are hence relevant to claims to proteins encoded by the DNA and antibodies which specifically bind to the proteins) commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

> If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

> In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

> 1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

> A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides and polypeptides in terms of chemical structure, rather than functional characteristics. For example, the "variant language" of independent claim 1 recites chemical structure to define the claimed genus:

> 1. An isolated cDNA comprising a nucleic acid encoding an amino acid sequence selected from:...c) a variant of SEQ ID NO:1 having at least 90% amino acid sequence identity to SEQ ID NO:1...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the

present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the polynucleotides or polypeptides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polynucleotides or polypeptides or. The polynucleotides or polypeptides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*

2. The present claims do not define a genus which is "highly variant"

Furthermore, the claims at issue do not describe a genus which could be characterized as highly variant, i.e., "encompassing a substantial variety of subgenera" (Final Office Action, page 8). Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078; cited at page 29 of the instant application). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to GPCR proteins, in particular, metabotropic glutamate GPCR proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as metabotropic glutamate GPCR proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:1. The "variant language" of the present claims recites, for example, polynucleotides encoding "an amino acid sequence having at least 90% amino acid sequence identity SEQ ID NO:1" (note that SEQ ID NO:1 has 441 amino acid residues). This

variation is far less than that of all potential metabotropic glutamate GPCR proteins related to SEQ ID NO:1, i.e., those metabotropic glutamate GPCR proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:1.

3. The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of September 17, 1998. Much has happened in the development of recombinant DNA technology in the 20 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1 and SEQ ID NO:7, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants at the time of filing of this application.

4. Summary

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1 or SEQ ID NO:7. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polynucleotides or polypeptides defined by the present claims is adequately described, as evidenced by Brenner et al and

consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

Claims 1 and 3-6 stand rejected under 35 U.S.C. § 102(b) as anticipated by Valenzuela et al. (WO 99/55271, November 4, 199) and, alternatively under 35 U.S.C. § 102(e) as anticipated by Moore et al. (U.S. Published Application 2003005536, effective filing date June 17, 1999). The rejection alleges in particular, that:

- Valenzuela disclose a nucleic acid molecule (SEQ ID NO:43, claim 52) that encodes a protein (SEQ ID NO:45, claim 53) that is 100% identical to the polypeptide of SEQ ID NO:7 of the instant application, thus anticipating the claims. Valeuzuela et al. also teach vectors, host cells, a method of producing protein, and labeled cDNA.

- Moore et al. disclose a nucleic acid moecule (SEQ ID NO:22) that encodes a protein (SEQ ID NO:146) that is 100% identical to the polypeptide of SEQ ID NO:1 from amino acids 1-384 of the instant application, and therefore discloses an isolated cDNA encoding a fragment of SEQ ID NO:1 from I51-V72, G88-V109, C116-A145, I156-L175, M207-P229, or G242-T264 of SEQ ID NO:1, as recited in claim 1. Moore et al also teach vectors, host cells, and a method of making a protein, therefore anticipating claims 3-6 as well.

- Because the instant application does not meet the requirements of 35 U.S.C. § 112, first paragraph, for the reasons given above, and it is a continuation of application Serial No. 09/516,513, the prior application does not meet these requirements and therefore is unavailable under 35 U.S.C. § 120. Under these circumstances, Valenzuela et al. and Moore et al. anticipate the claimed invention.

The now claimed invention, at least as recited in claims 1 and 3-6, is supported by both a specific and substantial asserted utility and a well established utility that is disclosed and enabled in priority application Serial No. 09/516,513

Applicants submit that, for the reasons cited above in response to the rejection of claims under 35 U.S.C. §§ 101/112, the specification supports a specific and substantial asserted utility, as well as a well established utility for the claimed invention that is similarly disclosed in the priority application Serial No. 09/516,513 in accordance with 35 U.S.C. § 120, therefore providing an effective filing date for the instant application of September 17, 1998.
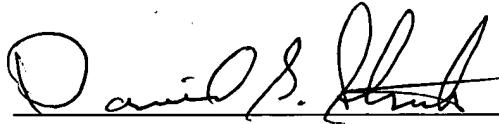
Due to the urgency of this matter and its economic and public health implications, an expedited review of this appeal is earnestly solicited.

If the USPTO determines that any additional fees are due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108.**

This brief is enclosed in triplicate.

Respectfully submitted,

INCYTE CORPORATION

Date: ___January 7, 2004___

David G. Streeter, Ph.D.
Reg. No. 43,168
Direct Dial Telephone: (650) 845-5741

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

Attachments:

1) Lashkari et al., Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR, Proc. Nat. Acad. Sci. 94:8945-8947

2) Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999);

3) Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000).

4) John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, 107 Environ. Health Perspec. 107:681-685 (1999).

5) Email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.

6) Page, M.J. et al., <u>Proteomics: a major new technology for the drug discovery process</u>, Drug Discov. Today 4:55-62 (1999).

7) Declaration of Tod Bedilion, Ph.D., under 37 C.F.R. 1.132;

8) Declaration of Vishwanath R. Iyer, Ph.D., under 37 C.F.R. 1.132 with Exhibits A - E; and

9) ten (10) references published before the filing date of the instant application:

> a) WO 95/21944, SmithKline Beecham, "Differentially expressed genes in healthy and diseased subjects" (Aug. 17, 1995)
>
> b) WO 95/20681, Incyte Pharmaceuticals, "Comparative Gene Transcript Analysis" (Aug 3, 1995)
>
> c) Schena et al., "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," Science 270:467-470 (Oct 20, 1995)
>
> d) WO 95/35505, Stanford University, "Method and apparatus for fabricating microarrays of biological samples" (Dec 28, 1995)
>
> e) U.S. Pat. No. 5,569,588, Ashby et al., "Methods for Drug Screening" (Oct 29, 1996)
>
> f) Heller al., "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays," *PNAS* 94:2150 - 2155 (Mar 1997)
>
> g) WO 97/13877, Lynx Therapeutics, "Measurement of Gene Expression Profiles in Toxicity Determinations" (April 17, 1997)
>
> h) Acacia Biosciences Press Release (August 11, 1997)
>
> i) Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," Genetic Engineering News (Sept. 15, 1997)
>
> j) DeRisi *et al.,* "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science 278:680 - 686 (Oct 24, 1997)

<u>APPENDIX - CLAIMS ON APPEAL</u>

1.   An isolated cDNA comprising a nucleic acid encoding an amino acid sequence selected from:

   a) an amino acid sequence of SEQ ID NO:1;

   b) a fragment of SEQ ID NO:1 from I51-V72, G88-V109, C116-A145, I156-L175, M207-P229, or G242-T264 of SEQ ID NO:1;

   c) a variant of SEQ ID NO:1 having at least 90% amino acid sequence identity to SEQ ID NO:1; and

   d) the complement of the encoding nucleic acid sequence of a), b), or c).

2.   An isolated cDNA comprising a nucleic acid sequence selected from:

   a) SEQ ID NO:7; and

   b) a variant of SEQ ID NO:7 having at least 95% identity to SEQ ID NO:7.

3.   A composition comprising the cDNA of claim 1 and a labeling moiety.

4.   A vector comprising the cDNA of claim 1.

5.   A host cell comprising the vector of claim 4.

6.   A method for using a cDNA to produce a protein, the method comprising:

   a)   culturing the host cell of claim 5 under conditions for protein expression; and

   b)   recovering the protein from the host cell culture.

Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI*†, JOHN H. MCCUSKER‡, AND RONALD W. DAVIS*§

*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and ‡Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

ABSTRACT   The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of Saccharomyces cerevisiae, Escherichia coli, Haemophilus influenzae (1), Mycoplasma genitalium (2), and Methanococcus jannaschii (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode Caenorhabditis elegans (4) and the small flowering plant Arabidopsis thaliana (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in S. cerevisiae have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as S. cerevisiae, will be extremely important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5' end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

†Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.
§To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu.
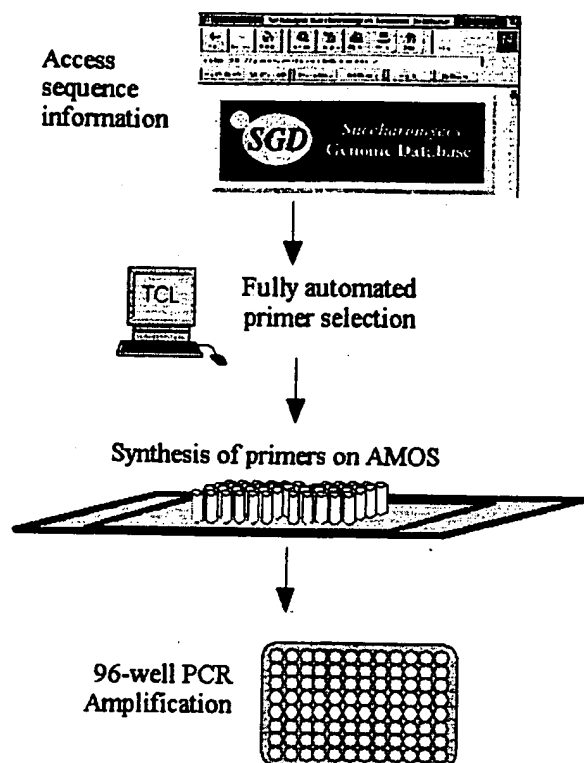
Fig. 1. Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.
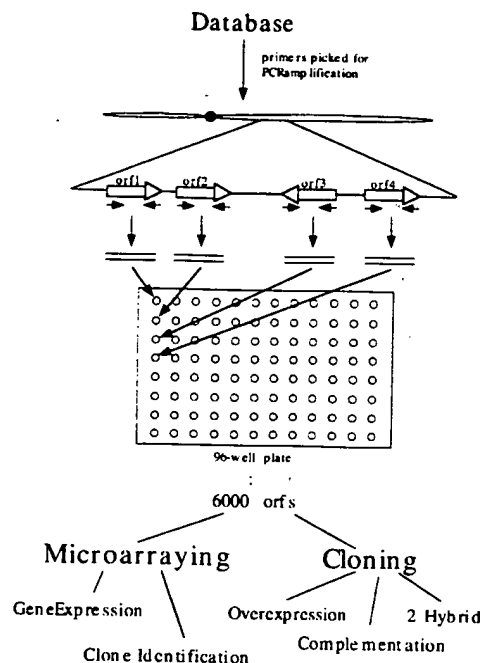

Fig. 2. Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose.

Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a "snapshot" of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

Applied Biological Sciences: Lashkari *et al.*

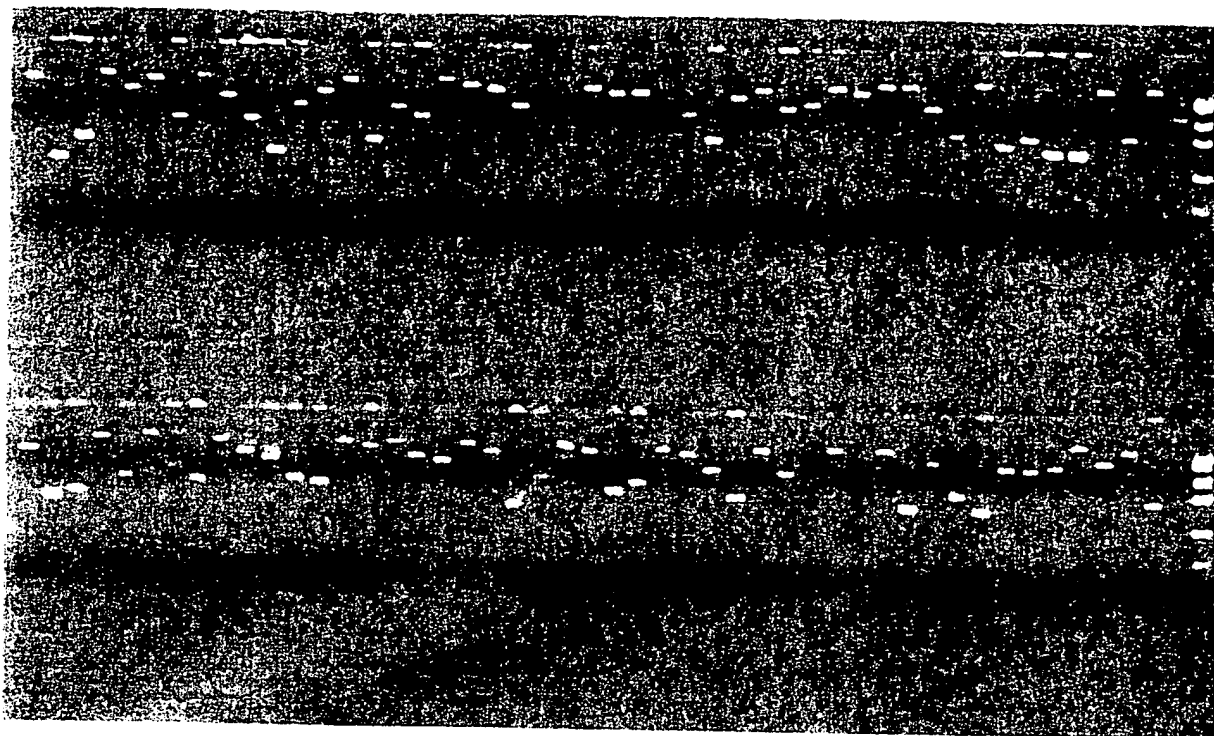*Proc. Natl. Acad. Sci. USA* 94 *(1997)*   8947

Fig. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct readout. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* 269, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* 270, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* 273, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* 356, 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* 106, 1241–1255.
6. Oliver, S. (1996) *Nature (London)* 379, 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* 92, 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* 57, A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* 270, 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* 251, 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* 274, 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* 14, 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* 274, 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* 316, 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* 89, 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* 80, 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* 340, 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1994) *Nat. Genet.* 4, 11–18.

■ IN PERSPECTIVE ■

Claudio J. C nti, Editor

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1]Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina
[2]Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. *Mol. Carcinog. 24:153–159, 1999.* © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, *Haemophilus influenzae,* was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, *Saccharomyces cervisiae* [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the *Homo sapiens* DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the BRCA1 gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on relative fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.
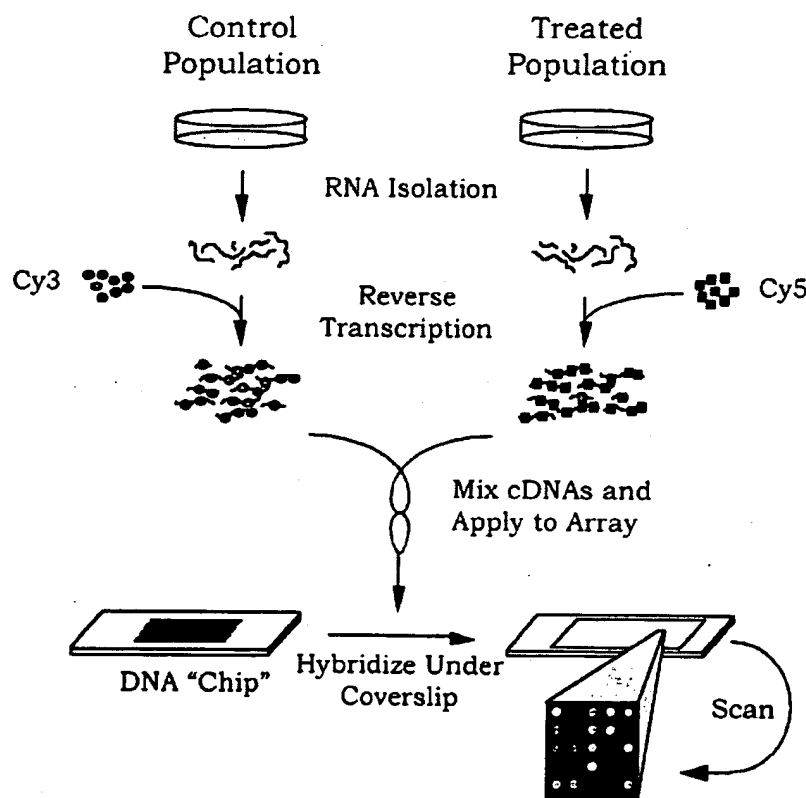

Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.
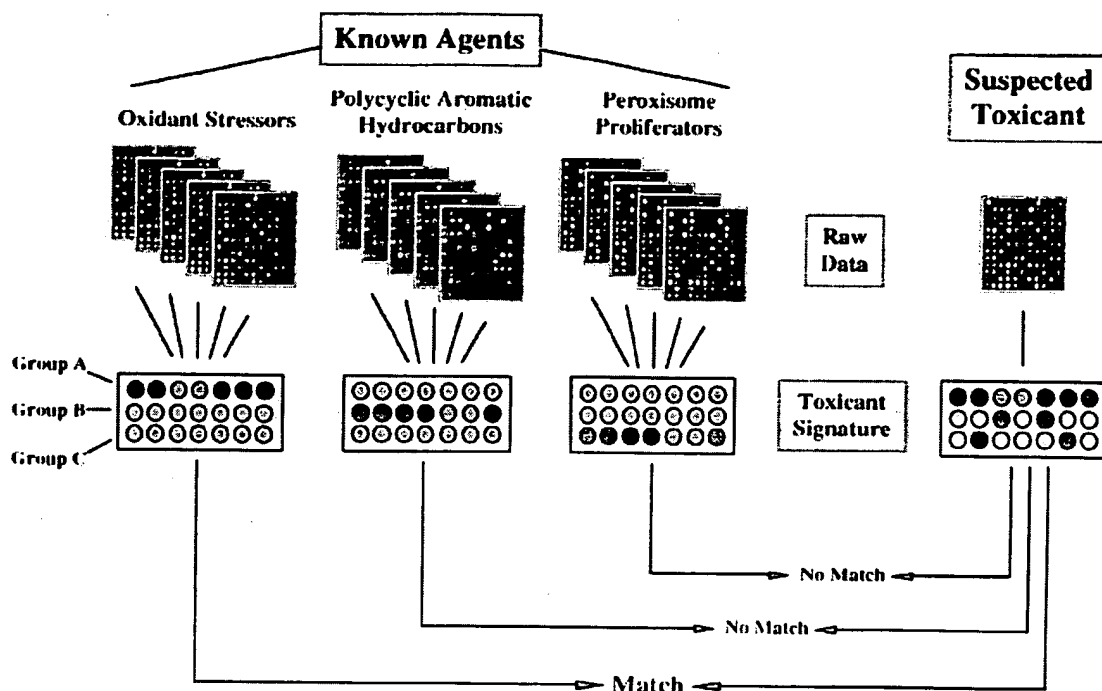
Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing Tox-Chip v2.0 and chips for other model systems, including rat, mouse, *Xenopus*, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip D sign d to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia,* and *Arabadopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.

20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.

21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.

22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.

23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.

24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.

25. http://www.resgen.com/

26. http://www.genomesystems.com/

27. http://www.clontech.com/

28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.

29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.

30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.

31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.

32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.

33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.

34. http://www.mdyn.com/

35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.

36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.

37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.

38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.

39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.

40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.

41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.

42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.

43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.

44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.

45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html

46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.

47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.

48. http://www.niehs.nih.gov/envgenom/home.html

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

* Corresponding author. Tel.: + 1-301-4245989; fax: + 1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality

Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.

F344 Rat Liver

Molecular Weight (kDaltons)

Isoelectric point (pH units)

Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al., 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

References

Aicher, L., Wahl, D., Arce, A., Grenet, O., Steiner, S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998–2003.

Anderson, N.L., Seilhamer, J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18, 533–537.

Anderson, N.L., Esquer-Blasco, R., Hofmann, J.P., Anderson, N.G., 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12, 907–930.

Anderson, L., Steele, V.K., Kelloff, G.J., Sharma, S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22, 108–116.

Anderson, N.L., Esquer-Blasco, R., Richardson, F., Foxworthy, P., Eacho, P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137, 75–89.

Arce, A., Aicher, L., Wahl, D., Esquer-Blasco, R., Anderson, N.L., Cordier, A., Steiner, S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63, 2243–2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepsdorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expresssion patterns with a complementary DNA microarray. Science 251, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Rayrnackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. *Key words:* DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681–685 (1999). [Online 6 July 1999]
http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (*1*). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer

has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

Array El ments

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in

a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (*2*). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (3). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of > 2,500 spots/cm^2 may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental

arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors effecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., ^{33}P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosporimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (4). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA$^+$ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After

Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (14), with permission from Genetic Microsystems.

normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5' end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain > 10^8 molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (5). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (6,7). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, *C.*

Table 1. Advantages and disadvantages of different microarray scanning systems.

		Nonconfocal laser scanner	
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples		May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device.
From Kawasaki (13).

elegans knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes

transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (8) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic

significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

• Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.
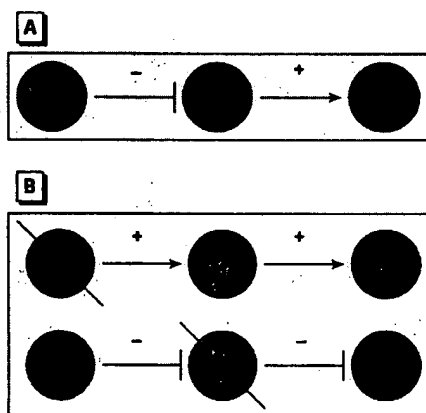
Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2. (A) A simple, two-component, linear regulatory network operating on gene i_2, where i_1 is a positive effector of i_2 and j_n is either a positive or negative effector of i_1. This network could be deduced by examining the consequence of (B) deleting j_n on the expression of i_1 and i_2, where the expression of i_2 would be decreased or increased depending on whether j_n was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15).

• Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, noncontaminated, sequence-verified, species/cell/tissue/field-specific clones.

• Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.

• Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.

• Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.

• Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).

• Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings.

Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.

• Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.

• Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.

• Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (9), and how this may compromise sperm counts and quality following sexual maturation (10). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (11) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene X is related to the expression of gene Y, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach

is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (12):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

REFERENCES AND NOTES

1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March 31:56–73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

SPEAKERS

Cindy Afshari
NIEHS
Linda Birnbaum
U.S. EPA
Ron Butow
University of Texas Southwestern Medical Center
Alex Chenchik
Clontech Laboratories, Inc.
David Dix
U.S. EPA

Abdel Elkahloun
Research Genetics, Inc.
Sue Fenton
U.S. EPA
Norman Hecht
University of Pennsylvania
Pat Hurban
Paradigm Genetics, Inc.
Bob Kavlock
U.S. EPA
Ernie Kawasaki
General Scanning, Inc.

Steve Krawetz
Wayne State University
Nick Mace
Genetic Microsystems, Inc.
Scott Mordecai
Affymetrix, Inc.
Kevin Morgan
Glaxo Wellcome, Inc.
Elaine Poplin
Research Genetics, Inc.
Don Rose
Cartesian Technologies, Inc.

Jim Samet
U.S. EPA
Sam Ward
University of Arizona
Jeff Welch
U.S. EPA
Reen Wu
University of California at Davis
Tim Zacharewski
Michigan State University

Subject: RE: [Fwd: Toxicology Chip]
   Date: Mon, 3 Jul 2000 08:09:45 -0400
   From: "Afshari,Cynthia" <afshari@niehs.nih.gov>
      To: "Diana Hamlet-Cox" <dianahc@incyte.com>

You can see the list of clones that we have on our 12K chip at
http://manuel.niehs.nih.gov/maps/guest/clonesrch.cfm
We selected a subset of genes (2000K) that we believed critical to tox
response and basic cellular processes and added a set of clones and ESTs to
this. We have included a set of control genes (80-) that were selected by
the NHGRI because they did not change across a large set of array
experiments. However, we have found that some of these genes change
significantly after tox treatments and are in the process of looking at the
variation of each of these 80- genes across our experiments.
Our chips are constantly changing and being updated and we hope that our
data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> From: Diana Hamlet-Cox
> Sent: Monday, June 26, 2000 8:52 PM
> To: afshari@niehs.nih.gov
> Subject: [Fwd: Toxicology Chip]
>
> Dear Dr. Afshari,
>
> Since I have not yet had a response from Bill Grigg, perhaps he was not
> the right person to contact.
>
> Can you help me in this matter? I don't need to know the sequences,
> necessarily, but I would like very much to know what types of sequences
> are being used, e.g., GPCRs (more specific?), ion channels, etc.
>
> Diana Hamlet-Cox
>
> -------- Original Message --------
> Subject: Toxicology Chip
> Date: Mon, 19 Jun 2000 18:31:48 -0700
> From: Diana Hamlet-Cox <dianahc@incyte.com>
> Organization: Incyte Pharmaceuticals
> To: grigg@niehs.nih.gov
>
> Dear Colleague:
>
> I am doing literature research on the use of expressed genes as
> pharmacotoxicology markers, and found the Press Release dated February
> 29, 2000 regarding the work of the NIEHS in this area. I would like to
> know if there is a resource I can access (or you could provide?) that
> would give me a list of the 12,000 genes that are on your Human ToxChip
> Microarray. In particular, I am interested in the criteria used to
> select sequences for the ToxChip, including any control sequences
> included in the microarray.
>
> Thank you for your assistance in this request.

07/31/2000 10:34 AM

Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.

Among the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should produce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic[1,2] and microarray[3,4] technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

Martin J. Page*, Bob Amess, Christian Rohlff, Colin Stubberfield and **Raj Parekh**, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. *tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk

Figure 1. Steps involved in analysing a biological sample by proteomics. MCI, molecular cluster index.
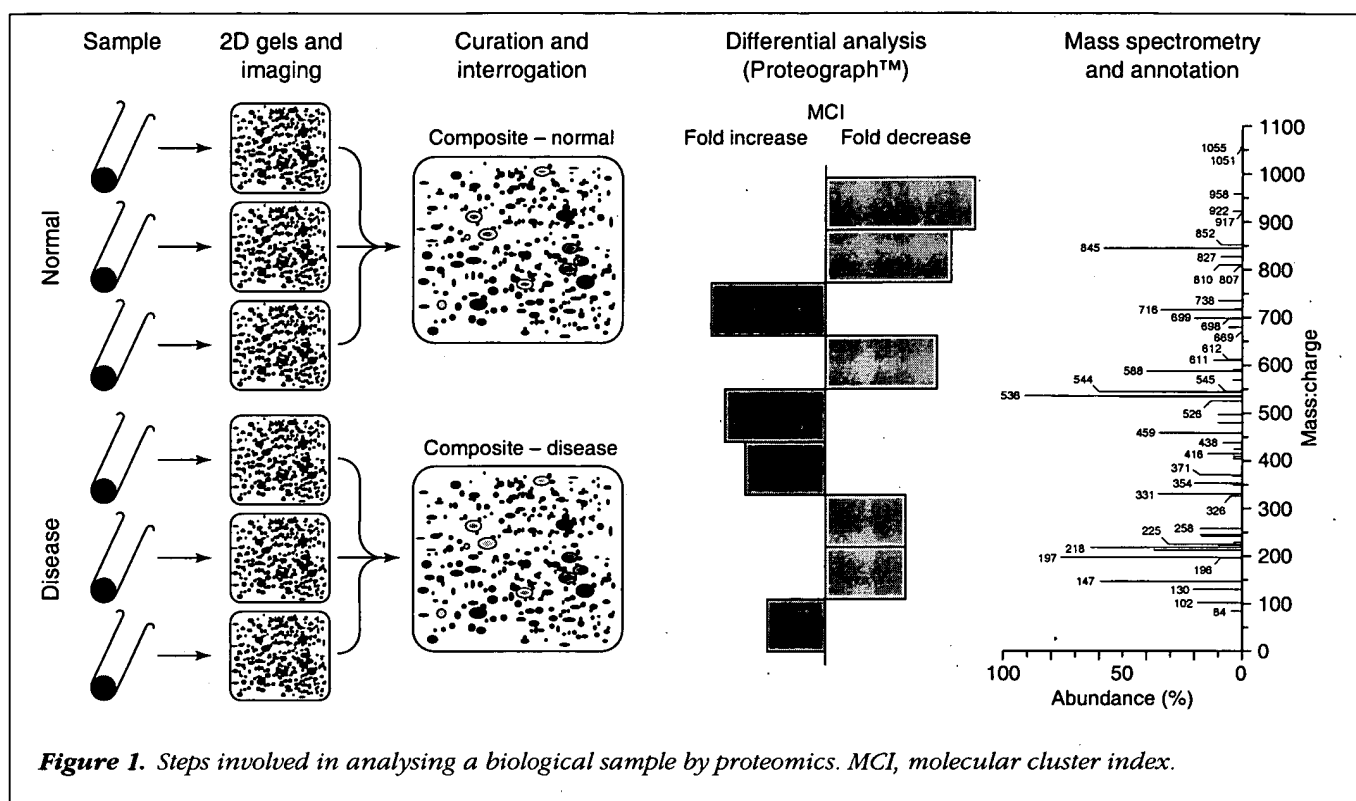
analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

Proteomics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed[5-7]. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins[8], which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the

process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS–PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one

another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

Use f proteomics to identify
disease specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and im-

mediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical

Figure 2. Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus Candida albicans and (c) the human hepatoma cell line Huh7.

Foregrounds: Huh7 cells treated with 5FU
Backgrounds: Huh7 cells untreated
■■■■■■■■   Upregulated in Huh7 cells treated with 5FU
with respect to untreated Huh7 cells
■■■■■■■■   Downregulated in Huh7 cells treated with 5FU
with respect to untreated Huh7 cells

MCI	Fold change		pI	MW
3589	■■■■■■ 16.3		5.60	13718
4035	■■■■ 15.8		8.73	13224
4033	■■■■ 12.3		8.94	13117
4052	■■■■ 11.7		8.96	21434
4064	■■■■ 11.1		9.37	29389
3792	■■■■ 10.9		8.75	41162
3976	■■■■ 10.5		7.14	12061
3672	■■■■ 10.5		8.83	24121
3986	■■■■ 10.4		4.61	10853
3759	■■■■ 9.9		4.98	41420
4032	■■■■ 9.7		4.56	13117
1972		−9.4 ■■■■	7.92	12396
3587	■■■■ 9.4		4.50	14954
2033		−9.4 ■■■■	7.93	11126
3984	■■■■ 8.9		5.32	11090
2403		−8.8 ■■■■	6.04	25950
3748	■■■■ 8.8		6.32	35513
2105		−8.7 ■■■■	4.76	20803
3897	■■■■ 8.7		4.94	87842
4221	■■■■ 8.6		5.78	71963

Figure 3. Table of differential protein expression profiles, referred to as a Rosetta Proteograph™, between Huh7 cells with and without the cytotoxic agent 5-FU. Bars are quantized and do not represent exact fold change values.

mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry[9]. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas[10], human breast proteins from normal and tumour sources[11–13], lung tumours[14], colon tumours[15] and bladder tumours[16]. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified[17,18].

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. et al., submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput

Pr te mics for targ t validati n and signal transduc-ti n studies

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences[20].

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules[21-23]. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

Immunoprecipitation studies

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics[24,25]. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affin-ity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies[26-28]. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1–10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable

Figure 4. Features that are specifically up- or downregulated in Huh7 cells by either 5-fluorouracil (5-FU) or OGT719: (a) elongation factor 1α2, (b) novel (three peptides by MS-MS) and (c) α-subunit of prolyl-4-hydroxylase. Arrows indicate up- or downregulated.

of targeting, and being retained in, cells bearing the asialo-glycoprotein receptor (ASGP-r), including hepatocytes[29], hepatoma Huh7 cells[30] and some colorectal tumour cells[31]. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with IC_{50} doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins.

Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein[32], can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

Clear potential

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

Use of proteomics in formal drug t xicology studies

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from in vitro studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems in vivo. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an in vivo setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members[33,34], encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabo-lism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

Unique P450 profiles

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up[35–37]. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

C nclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

REFERENCES

1 Crooke, S.T. (1998) Nat. Biotechnol. 16, 29–30
2 Dykes, C.W. (1996) Br. J. Clin. Pharmacol. 42, 683–695
3 Schena, M. et al. (1998) Trends Biotechnol. 16, 301–306
4 Ramsay, G. (1998) Nat. Biotechnol. 16, 40–44
5 Anderson, N.L. and Anderson, N.G. (1998) Electrophoresis 19, 1853–1861
6 James, P. (1997) Biochem. Biophys. Res. Commun. 231, 1–6
7 Wilkins, M.R. et al. (1996) Biotechnol. Genet. Eng. Rev. 13, 19–50
8 Parekh, R.B. and Rohlff, C. (1997) Curr. Opin. Biotechnol. 8, 718–723
9 Figeys, D. et al. (1998) Electrophoresis 19, 1811–1818
10 Wimmer, K. et al. (1996) Electrophoresis 17, 1741–1751
11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) Electrophoresis 18, 573–581
12 Williams, K. et al. (1998) Electrophoresis 19, 333–343
13 Rasmussen, R.K. et al. (1998) Electrophoresis 19, 818–825
14 Hirano, T. et al. (1995) Br. J. Cancer 72, 840–848
15 Ji, H. et al. (1997) Electrophoresis 18, 605–613
16 Ostergaard, M. et al. (1997) Cancer Res. 57, 4111–4117
17 Patel, V.B. et al. (1997) Electrophoresis 18, 2788–2794
18 Arnott, D. et al. (1998) Anal. Biochem. 258, 1–18
19 Anderson, L. and Seilhamer, J. (1997) Electrophoresis 18, 533–537
20 Rastan, S. and Beeley, L.J. (1997) Curr. Opin. Genet. Dev. 7, 777–783
21 Gravel, P. et al. (1995) Electrophoresis 16, 1152–1159
22 Qian, Y. et al. (1997) Clin. Chem. 43, 352–359
23 Sanchez, J.C. et al. (1997) Electrophoresis 18, 638–641
24 Watts, A.D. et al. (1997) Electrophoresis 18, 1086–1091
25 Asker, N. et al. (1995) Biochem. J. 308, 873–880
26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) Electrophoresis 15, 265–277
27 Huber, L.A. (1995) FEBS Lett. 369, 122–125
28 Corthals, G.L. et al. (1997) Electrophoresis 18, 317–323
29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) J. Cell Biol. 96, 217–229
30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) Biochem. Biophys. Res. Commun. 218, 325–330
31 Mu, J-Z. et al. (1994) Biochim. Biophys. Acta 1222, 483–491
32 Ghoshal, K. and Jacob, S.T. (1997) Biochem. Pharmacol. 53, 1569–1575
33 Guengerich, F.P. and Parikh, A. (1997) Curr. Opin. Biotechnol. 8, 623–628
34 Rendic, S. and Di Carlo, F.J. (1997) Drug Metab. Rev. 29, 413–580
35 Vermes, A., Guchelaar, H.J. and Koopmans, R.P. (1997) Cancer Treat. Rev. 23, 321–339
36 Housman, D. and Ledley, F.D. (1998) Nat. Biotechnol. 16, 492–493
37 Persidis, A. (1998) Nat. Biotechnol. 16, 209–210

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF TOD BEDILION, Ph.D.
UNDER 37 C.F.R. § 1.132

I, TOD BEDILION, Ph.D., declare and state as follows:

1.   In April, 1996, I became the first employee of Synteni, Inc., where I served as Research Director until its acquisition by Incyte Corporation in early 1998. After Synteni's acquisition, I continued in the position of Director of Corporate Development at Incyte until May 11, 2001. I am currently the Director of Business Development at Genomic Health, Inc., Redwood City, California and an occasional Consultant to Incyte.

2.   Synteni was founded to commercialize expression microarrays, microarrays in which expressed nucleic acids -- full-length cDNAs, fragments of full-length cDNAs, expressed sequence tags (ESTs) -- are arrayed on a common support to permit highly parallel detection and measurement of the expression of their cognate genes in a biological sample.

3.   During my employ at Synteni, virtually all (if not all) of my work efforts were directed to the further technical development and the commercial exploitation of that microarray technology; given the small size of our shop, most of us had both technical and commercial responsibilities. The customer accounts for which I was personally responsible included large pharmaceutical companies, such as SmithKline

Beecham, large biotechnology companies, such as Genentech, and small research institutes, such as DNAX Inc.

4.   From my very first interaction with our customers, consistently through to Synteni's acquisition by Incyte, I heard uniform, consistent, and emphatic requests that more genes be added to the arrays. This was true with respect to both our original microarrays, based on customer-provided genes and libraries, and our later, "generic", gene expression microarrays, based upon the unigene clone collection (our so-called "UniGem" arrays). From day 1, the pressure on us was to print ever more spots on the array. It was **never** a question: our customers wanted ever more genes on the array, each new gene-specific probe providing incrementally more value to the customer.[1]

5.   As a commercial enterprise, providing value to our customers was our major concern. Thus, to increase the value of our products and services in the marketplace -- to increase our ability to sell our microarrays and microarray services, their "salability" -- our efforts from the very beginning were devoted to increasing the number of specific genes whose expression could be detected with our microarrays.

6.   Indeed, one of our major competitive advantages in the marketplace -- not just as regards other commercial suppliers, but also with respect to the innumerable laboratories and companies that were attempting to spot arrays in their own "home-brew" facilities -- was the number of

---

[1]   I should note the customers were not asking for addition of probes specific to only those genes for which the biological function of the encoded gene product was known, but were asking for probes specific to any and all expressed genes.

distinct gene-specific probes that we provided on our expression microarrays. Our first 10,000 element UniGem array put the holy grail of gene expression analysis -- the human whole genome array -- within sight for the very first time (with respect to timing of the UniGEM program we began project planning and technology development in mid 1996 and delivered our first 10,000 element standard content human arrays in the first months of 1997 as I recall).

7. By the end of 1997, our efforts to provide the most comprehensive, and thus most valuable, human gene expression microarrays had been sufficiently successful that Incyte agreed to acquire Synteni for a reported $80 million.

8. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.

Tod Bedilion, Ph.D.   10/20/03
                                       Date

-3-

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF VISHWANATH R. IYER, Ph.D.
UNDER 37 C.F.R. § 1.132

I, VISHWANATH R. IYER, Ph.D., declare and state as follows:

1. I am an Assistant Professor in the Section of Molecular Genetics and Microbiology, Institute of Cellular and Molecular Biology, University of Texas at Austin, where my laboratory currently studies global transcriptional control in yeast, gene expression programs during human cell proliferation, and genome-wide transcription factor targets in yeast and human. Immediately prior to this position, I spent four years as a postdoctoral fellow in the laboratory of Patrick O. Brown at Stanford University studying the transcriptional programs of yeast and of human cells. My *curriculum vitae* is attached hereto as Exhibit A.

2. Beginning in Dr. Brown's laboratory, where I helped to develop the first whole genome arrays for yeast and early versions of highly representative cDNA arrays for human cells, and continuing to the present day, I have used microarray-based gene expression analysis as a principal approach in much of my research.

3. Representative publications describing this work include:

DeRisi J. et al., "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278:680-686 (1997);[1]

Marton et al., "Drug target validation and identification of secondary drug target effects using DNA microarrays," Nature Med. 4:1293-1301 (1998);[2]

Iyer et al., "The transcriptional program in the response of human fibroblasts to serum," Science 283:83-87 (1999);[3] and

Ross et al., "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics* 24: 227-235 (2000).[4]

Two of the papers describe our use of microarray-based expression profiling to explore the metabolic reprogramming that occurs during major environmental changes, both in yeast (DeRisi et al., during the shift from fermentation to respiration) and in human cells (Iyer et al., human fibroblasts exposed to serum). One reference describes our use of expression profile analysis in drug target validation and identification of secondary drug effects (Marton et al.). And one describes our use of expression profiling as a molecular phenotyping tool to discriminate among human cancer cells (Ross et al.).

4. Whether used to elucidate basic physiological responses, to study primary and secondary drug effects, or to discriminate and classify human cancers, expression profiling

---

[1] Attached hereto as Exhibit B.

[2] Attached hereto as Exhibit C.

[3] Attached hereto as Exhibit D.

[4] Attached hereto as Exhibit E.

as we have practiced it relies for its power on comparison of **_patterns_** of expression.

5.   For example, we have demonstrated that we can use the presence or absence of a characteristic drug "signature" pattern of altered gene expression in drug-treated cells to explore the mechanism of drug action, and to identify secondary effects that can signal potentially deleterious drug side effects. As another example, we have demonstrated that gene expression patterns can be used to classify human tumor cell lines. While it is of course advantageous to know the biological function of the encoded gene products in order to reach a better understanding of the cellular mechanisms underlying these results, these pattern-based analyses do not require knowledge of the biological function of the encoded proteins.

6.   The resolution of the patterns used in such comparisons is determined by the number of genes detected: the greater the number of genes detected, the higher the resolution of the pattern. It goes without saying that higher resolution patterns are generally more useful in such comparisons than lower resolution patterns. With such higher resolutions comes a correspondingly higher degree of statistical confidence for distinguishing different patterns, as well as identifying similar ones.

7.   Each gene included as a probe on a microarray provides a signal that is specific to the cognate transcript, at least to a first approximation.[5] Each new gene-specific

---

[5]   In a more nuanced view, it is certainly possible for a probe to signal the presence of a variety of splice variants of a single gene,

(Continued...)

-3-

probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device. As I note above, higher resolution patterns are generally more useful in comparisons than lower resolution patterns. Accordingly, each new gene probe added to a microarray increases the usefulness of the device in gene expression profiling analyses. This proposition is so well-established as to be virtually an axiom in the art, and has been as long as I have been working in the field, and certainly since the time I embarked on the production of whole genome arrays in early 1996. Simply put, arrays with fewer gene-specific probes are inferior to arrays with more gene-specific probes.

8. For example, our ability to subdivide cancers into discriminable classes by expression profiling is limited by the resolution of the patterns produced. With more genes contributing to the expression patterns, we can potentially draw finer distinctions among the patterns, thus subdividing otherwise indistinguishable cancers into a greater number of classes; the greater the number of classes, the greater the likelihood that the cancers classified together will respond similarly to therapeutic intervention, permitting better individualization of therapy and, we hope, better treatment outcomes.

9. If a gene does not change expression in an experiment, or if a gene is not expressed and produces no

_____

(_Continued)
without discriminating among them, and for a probe to signal the presence of a variety of allelic variants of a single gene, again without discriminating among them.

signal in an experiment, that is not to say that the probe lacks usefulness on the array; it only means that an insufficient number of conditions have been sampled to identify expression changes. In fact, an experiment showing that a gene is not expressed or that its expression level does not change can be equally informative. To provide maximum versatility as a research tool, the microarray should include -- and as a biologist I would want my microarray to include -- each newly identified gene as a probe.

10. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.

_____
VISHWANATH R. IYER, Ph.D.

<u>October 20, 2003</u>
Date

Vishwanath R. Iyer

Assistant Professor
Section of Molecular Genetics and Microbiology
Institute of Cellular and Molecular Biology
MBB 3.212A, University of Texas at Austin
Austin, TX 78712-0159
Phone: 512-232-7833
Fax: 512-232-3432
Email: vishy@mail.utexas.edu

Education/Training

Bombay University, Mumbai, India	B.Sc. (1987), Chemistry & Biochemistry
M. S. University of Baroda, Baroda, India	M.Sc. (1989), Biotechnology
Harvard University, Cambridge MA	Ph.D. (1996), Genetics
Stanford University, Stanford CA	Post-doctoral (1996-2000), Genomics

Research Experience

9/00-5/03   Assistant professor, Section of Molecular Genetics and Microbiology, University of Texas, Austin TX
- Global transcriptional control in yeast
- Gene expression programs during human cell proliferation
- Genome-wide transcription factor targets in yeast and human
- Collaborative microarray facility

5/96-8/00   Post-doctoral fellow   Stanford University, Stanford CA
(Advisor: Dr. Patrick O. Brown)
- Yeast whole-genome ORF and intergenic microarrays
- Human cDNA microarrays for expression profiling

9/89-4/96   Graduate student   Harvard University, Cambridge MA
(Advisor: Dr. Kevin Struhl)
- Yeast transcriptional regulation

Honours and Awards

Government of India Biotechnology Fellowship (1987-1989)
University Grants Commission Junior Research Fellowship (1989)
Stanford University/NHGRI Genome Training Grant (1996)

Invited Conference talks (selected)

Invited Lecturer, NEC-Princeton Lectures in Biophysics
    Princeton, NJ (June 1998)
Plenary Session Speaker, HGM '99 (HUGO Human Genome Meeting)
    Brisbane, Australia (April 1999)
Invited Speaker, Gordon Research Conference "Human Molecular Genetics"
    Newport, RI (August 2001)

Invited Speaker, Nature Genetics "Oncogenomics 2002" Conference
 Dublin, Ireland (May 2002)
Invited Speaker, "Pathology Bioinformatics" Symposium, University of Michigan,
 Ann Arbor, MI (November 2002)
Invited Speaker, "Systems Biology: Genomic Approaches to Transcriptional
 Regulation" Cold Spring Harbor Laboratory Meeting (March 2003)
Symposium co-Chair and Speaker "Functional Genomics" American Society for
 Biochemistry and Molecular Biology Meeting, San Diego, CA (April 2003)
Invited Speaker in Functional Genomics (Gene Networks) Symposium, International
 Congress of Genetics, Melbourne Australia July 6-11 2003
Invited Speaker "BioArrays Europe 2003"
 Cambridge, UK (Sep/Oct 2003)

Departmental Seminars
Texas A&M University Genetics and Biochemistry & Biophysics Departments,
 October 24 2002
New York University School of Medicine, Department of Biochemistry,
 November 20 2002
UT Southwestern Medical Center, Human Genetics Seminar Series,
 May 5 2002
UCLA School of Medicine, Department of Human Genetics
 June 2 2003
National Human Genome Research Institute
 June 12 2003
Sanger Institute of the Wellcome Trust, Hinxton, UK
 Sep 2003

Other Professional Activities
Reviewer for *Genome Biology, Genome Research, Nature Genetics, Science* (1998-
 2003)
Instructor, Cold Spring Harbor Summer Course "Making and using DNA Microarrays"
 (2000 - 2003)
Member, NIDDK Special Emphasis Review Panel ZDK1 (2001-2002)

Publications
1. Iyer V. & Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that
 stimulates transcription via its intrinsic DNA structure, *EMBO J.* 14: 2570-2579.

2. Iyer V. & Struhl, K. (1995) Mechanism of differential utilization of the his3 TR and TC
 TATA elements, *Mol. Cell. Biol.* 15: 7059-7066.

3. Iyer V. & Struhl K. (1996) Absolute mRNA levels and transcription initiation rates in
 Saccharomyces cerevisiae. Proc. Natl. Acad. Sci . (USA) 93:5208-5212.

4. DeRisi J. L., Iyer V. R. & Brown P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686

5. Marton M. J., DeRisi J. L., Bennett H. A., Iyer V. R., Meyer M. R., Roberts C. J., Stoughton R., Burchard J., Slade D., Dai H., Bassett D. E. Jr., Hartwell L. H., Brown P. O. & Friend S. H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* 4:1293-1301

6. Lutfiyya L. L., Iyer V. R., DeRisi J., DeVit M. J., Brown P. O. & Johnston M. (1998) Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae. Genetics* 150:1377-1391

7. Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. & Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273-3297

8. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J. C.,F., Trent J. M., Staudt L. M., Hudson Jr. J., Boguski M. S., Lashkari D., Shalon D., Botstein D. & Brown P. O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83-87

9. DeRisi J. L. & Iyer V. R. (1999) Genomics and array technology. *Curr. Opin. Oncol.* 11:76-79

10. Ross D. T., Scherf U., Eisen M. B., Perou C. M., Spellman P., Iyer V. R., Rees C., Jeffrey S. S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J. C. F., Lashkari D., Shalon D., Myers T. G., Weinstein J. N., Botstein D., & Brown P. O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24: 227-235

11. Sudarsanam P., Iyer V. R., Brown P. O. & Winston F. (2000) Whole-genome expression analysis of *snf/swi* mutants of *S. cerevisiae. Proc. Natl. Acad. Sci .(USA)* 97: 3364-3369

12. Tran H. G., Steger D. J., Iyer V. R., & Johnson A. D. (2000) The chromo domain protein Chd1p from budding yeast is an ATP-dependent chromatin-modifying factor *EMBO J* 19: 2323-2331

13. Gross C., Kelleher M., Iyer V. R., Brown P. O., & Winge D. R.. (2000) Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays. *J. Biol. Chem.* 275: 32310-32316

14. Reid J. L., Iyer V. R., Brown P. O. & Struhl K. (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* 6: 1297–1307

15. Iyer V. R., Horak C., Scafe C. S., Botstein D., Snyder M. & Brown P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF *Nature* 409: 533-538

16. Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J.L., Eisen M.B., Iyer V.R., Brown P.O., Muramatsu M., Shimada H., Okazaki Y. & Hayashizaki Y. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays *Proc. Natl. Acad. Sci. (USA)* 98: 2199-2204

17. Pollack J. R. & Iyer V.R. (2002) Characterizing the physical genome. *Nature Genetics* 32 suppl: 515-521

18. Iyer V. R. Microarray-based detection of DNA protein interactions: Chromatin Immunoprecipitation on Microarrays, in *DNA Microarrays: A Molecular Cloning Manual* (eds. Bowtell, D. & Sambrook, J.) 453-463 (Cold Spring Harbor Laboratory Press, 2003).
(not peer reviewed)

19. Killion, P., Sherlock G. and Iyer V. R. (2003) The Longhorn Array Database, an open-source implementation of the Stanford Microarray Database *BMC Bioinformatics* 4: 32

20. Hahn J. S., Hu Z., Thiele D. J. & Iyer V. R. Genome-Wide Analysis of the Biology of Stress Responses Through Heat Shock Transcription Factor (submitted to *PNAS*)

21. Kim J. & Iyer V.R. The global role of TBP recruitment to promoters in mediating gene expression profiles (manuscript in preparation)

Current/Pending Research Support

003658-0223-2001 Iyer (PI) 16% effort
01/01/02 - 08/31/04
Texas Higher Education Coordinating Board (ARP)
"Microarray based global mapping of DNA-protein interactions at promoters in human cells"
This is a pilot project to map the in vivo interactions of transcription factors with human promoters
Role: PI

Information Technology Research 0325116 R. Mooney (PI) 9% effort
09/01/03 - 08/31/07
NSF
"Feedback from Multi-Source Data Mining to Experimentation for Gene Network Discovery"
Role: Co-investigator

1 R01 CA95548-01A2 (pending) Iyer (PI) 25% effort
12/1/03 – 11/30/08
NIH
"Analysis of genome-wide transcriptional control in yeast"
This is a project to identify stress responsive transcription factor targets in yeast through the use of DNA microarrays
Role: PI

Breast Cancer Idea Award (pending) Iyer (PI) 10% effort
1/1/04 – 12/31/06
US Army Medical Research and Materiel Command
"Genome-wide chromosomal targets of oncogenic transcription factors"
This is a project aimed at identifying direct chromosomal targets of c-myc and ER in human cells through the use of a novel sequence tag analysis method.
Role: PI

003658-0531-2003 (pending) Marcotte (PI) 8% effort
01/01/04 - 12/31/05
Texas Higher Education Coordinating Board (ATP)
"Cell arrays: A novel high–throughput platform for measuring gene function on a genomic scale"
This proposal is aimed at developing a novel microarray based platform for automated, high–throughput microscopic imaging of cells, allowing rapid and systematic evaluation of gene function.

Fischer-Vize, Science 270, 1828 (1995).
35. T. C. James and S. C. Elgin, Mol. Cell Biol. 6, 3862 (1986); R. Paro and D. S. Hogness, Proc. Natl. Acad. Sci. U.S.A. 88, 263 (1991); B. Tschiersch et al., EMBO J. 13, 3822 (1994); M. T. Madireddi et al., Cell 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, Proc. Natl. Acad. Sci. U.S.A. 93, 7137 (1996).
36. P. M. Palosaari et al., J. Biol. Chem. 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E.

Grund, R. Eichenlaub, Appl. Environ. Microbiol. 58, 4068 (1992); V. Sharma, K. Suvama, R. Meganathan, M. E. Hudspeth, J. Bacteriol. 174, 5057 (1992); M. Kanazawa et al., Enzyme Protein 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, J. Bacteriol. 178, 3015 (1996).
37. M. Ho et al., Cell 77, 869 (1994).
38. W. Hendriks et al., J. Cell Biochem. 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of Saccharomyces cerevisiae were used to carry out a comprehensive investigation of the temporal program of gene expression acc mpanying the metabolic shift from fermentation to respiration. The expression profil s observed for genes with known metabolic functions pointed to features of the metab lic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The sam DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor TUP1 or overexpression of the transcriptional activator YAP1. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

Saccharomyces cerevisiae is an especially

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

t any gene whose function is kn wn (15). The responses of these previously uncharacterized genes t the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl–coenzyme A(CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes PCK1, encoding phosphoenolpyruvate carboxykinase, and FBP1, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were kn wn to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, ACR1 and IDP2, revealed that ACR1, a gene essential for ACS1 activity, also possessed a consensus CSRE motif, but interestingly, IDP2 did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception

Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of <5 × 10⁶ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of ~2 × 10⁸ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been shown to be controlled at least in part by these elements (21-24). Inspection of the sequences upstream of HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26-28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS$_{rpg}$) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of

downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the tup1Δ mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion f the TUP1 gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (40).

Wild-type yeast cells and cells bearing a deletion f the TUP1 gene (tup1Δ) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the tup1Δ strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the tup1Δ mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of TUP1, suggesting that these genes may be subject to TUP1-mediated repression by glucose. For example, SUC2, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of TUP1.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes MFA1 and MFA2, and the DNA damage–inducible RNR2 and RNR4, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of TUP1 itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the tup1Δ strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be TUP1-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when TUP1 was deleted. Another group of related genes that appeared to be subject to TUP1 repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the tup1Δ

Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by m re than sevenf ld when TUP1 was deleted. In contrast, none of 83 genes that could be classified as putative regulat rs of the cell division cycle were induced m re than tw f ld by deletion of TUP1. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall int a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which MFA1 and MFA2, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic tup1Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MATA strain (in which expression of MFA1 and MFA2 is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. YAP1 encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of YAP1 in yeast confers increased resistance to hydrogen peroxide, o-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing YAP1 under the control of the strong GAL1-10 promoter, both grown in galactose (that is, a condition that induces YAP1 overexpression). Complementary DNA from the control and YAP1 overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing YAP1.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in S. cerevisiae, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain verexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inosit l metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.

Table 1. Genes induced by YAP1 overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon YAP1 overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to A. thaliana zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found nly four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (**A**) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (**B**) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of *IDP2*, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (**C**) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (**D**) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (**E**) *SAM1*, *GPP1*, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (**F**) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, *Genome Res.* 6, 639 (1996).
3. D. Lashkari, *Proc. Natl. Acad. Sci. U.S.A.*, in press.
4. J. DeRisi et al., *Nature Genet.* 14, 457 (1996).
5. D. J. Lockhart et al., *Nature Biotechnol.* 14, 1675 (1996).
6. M. Chee et al., *Science* 274, 610 (1996).
7. M. Johnston and M. Carlson, in *The Molecular Biology of the Yeast Saccharomyces: Gene Expression*, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-
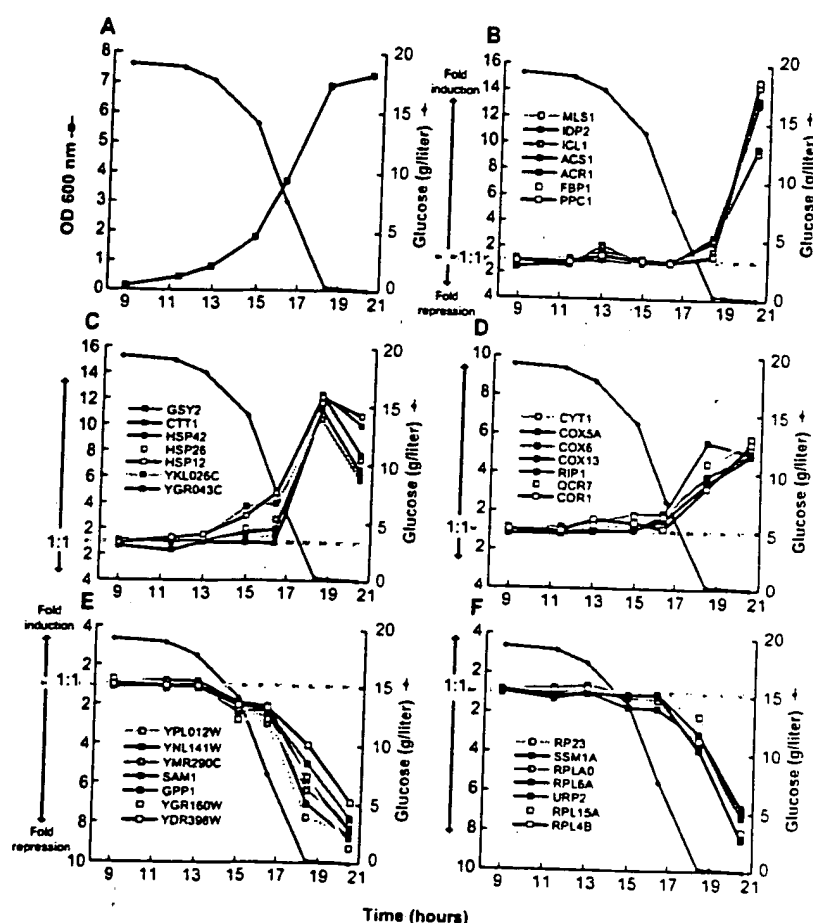
tion, the bound DNA was denatured by a 2-min incubation in distilled water at −95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.

11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)+] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).

12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then in 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.

13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html

14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.

15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx).

16. A. Scholer and H. J. Schuller, Mol. Cell. Biol. 14, 3613 (1994).

17. S. Kratzer and H. J. Schuller, Gene 161, 75 (1995).

18. R. J. Haselbeck and H. L. McAlister, J. Biol. Chem. 268, 12116 (1993).

19. M. Fernandez, E. Fernandez, R. Rodicio, Mol. Gen. Genet. 242, 727 (1994).

20. A. Hartig et al., Nucleic Acids Res. 20, 5677 (1992).

21. P. M. Martinez et al., EMBO J. 15, 2227 (1996).

22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, Mol. Cell. Biol. 15, 6232 (1995).

23. H. Ruis and C. Schuller, Bioessays 17, 959 (1995).

24. J. L. Parrou, M. A. Teste, J. Francois, Microbiology 143, 1891 (1997).

25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.

26. S. L. Forsburg and L. Guarente, Genes Dev. 3, 1166 (1989).

27. J. T. Olesen and L. Guarente, ibid. 4, 1714 (1990).

28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, Mol. Microbiol. 13, 119 (1994).

29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.

30. C. Fondrat and A. Kalogeropoulos, Comput. Appl. Biosci. 12, 363 (1996).

31. D. Shore, Trends Genet. 10, 408 (1994).

32. R. J. Planta and H. A. Raue, ibid. 4, 64 (1988).

33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.

34. S. F. Neuman, S. Bhattacharya, J. R. Broach, Mol. Cell. Biol. 15, 3187 (1995).

35. P. Lesage, X. Yang, M. Carlson, ibid. 16, 1921 (1996).

36. For example, we observed large inductions of the genes coding for PCK1, FBP1 [Z. Yin et al., Mol. Microbiol. 20, 751 (1996)], the central glyoxylate cycle gene ICL1 [A. Scholer and H. J. Schuller, Curr. Genet. 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, ACS1 [M. A. van den Berg et al., J. Biol. Chem. 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes PYK1 and PFK2 [P. A. Moore et al., Mol. Cell. Biol. 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T CTT1 [P. H. Bissinger et al., ibid. 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as HSP12, HSP26, and HSP42 [I. Farkas et al., J. Biol. Chem. 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, Mol. Gen. Genet. 223, 97 (1990); D. Wotton et al., J. Biol. Chem. 271, 2717 (1996)].

37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, FBP1 and PCK1) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).

38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, ADH1 and ADH2, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as HXK1/HXK2 (77% identical) [P. Herrero et al., Yeast 11, 137 (1995)], MLS1/DAL7 (73% identical) (20), and PGM1/PGM2 (72% identical) [D. Oh, J. E. Hopper, Mol. Cell. Biol. 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.

39. F. E. Williams, U. Varanasi, R. J. Trumbly, Mol. Cell. Biol. 11, 3307 (1991).

40. D. Tzamarias and K. Struhl, Nature 369, 758 (1994).

41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concor-

dance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.

42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of TUP1 and an Eco R I site 124 base pairs before the stop codon of the TUP1 gene.

43. L. R. Kowalski, K. Kondo, M. Inouye, Mol. Microbiol. 15, 341 (1995).

44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, Gene 148, 149 (1994).

45. D. Hirata, K. Yano, T. Miyakawa, Mol. Gen. Genet. 242, 250 (1994).

46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, Appl. Environ. Microbiol. 60, 1783 (1994).

47. A. Muheim et al., Eur. J. Biochem. 195, 369 (1991).

48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, J. Biol. Chem. 269, 32592 (1994).

49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm.stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.

50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)

51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Drug target validation and identification of secondary drug target effects using DNA microarrays

Matthew J. Marton[1], Joseph L. DeRisi[2], Holly A. Bennett[1], Vishwanath R. Iyer[2],
Michael R. Meyer[1], Christopher J. Roberts[1], Roland Stoughton[1], Julja Burchard[1],
David Slade[1], Hongyue Dai[1], Douglas E. Bassett, Jr[1]., Leland H. Hartwell[3],
Patrick O. Brown[2] & Stephen H. Friend[1]

[1]Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA
[2]Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute
Stanford, California 94305-5428, USA
[3]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109, USA
Correspondence should be addressed to S.H.F; email: sfriend@rosetta.org

We describe here a method for drug target validation and identification of secondary drug target effects based on genome-wide gene expression patterns. The method is demonstrated by several experiments, including treatment of yeast mutant strains defective in calcineurin, immunophilins or other genes with the immunosuppressants cyclosporin A or FK506. Presence or absence of the characteristic drug 'signature' pattern of altered gene expression in drug-treated cells with a mutation in the gene encoding a putative target established whether that target was required to generate the drug signature. Drug dependent effects were seen in 'targetless' cells, showing that FK506 affects additional pathways independent of calcineurin and the immunophilins. The described method permits the direct confirmation of drug targets and recognition of drug-dependent changes in gene expression that are modulated through pathways distinct from the drug's intended target. Such a method may prove useful in improving the efficiency of drug development programs.

Good drugs are potent and specific; that is, they must have strong effects on a specific biological pathway and minimal effects on all other pathways. Confirmation that a compound inhibits the intended target (drug target validation) and the identification of undesirable secondary effects are among the main challenges in developing new drugs. Comprehensive methods that enable researchers to determine which genes or activities are affected by a given drug might improve the efficiency of the drug discovery process by quickly identifying potential protein targets, or by accelerating the identification of compounds likely to be toxic. DNA microarray technology, which permits simultaneous measurement of the expression levels of thousands of genes, provides a comprehensive framework to determine how a compound affects cellular metabolism and regulation on a genomic scale[1-11]. DNA microarrays that contain essentially every open reading frame (ORF) in the *Saccharomyces cerevisiae* genome have already been used successfully to explore the changes in gene expression that accompany large changes in cellular metabolism or cell cycle progression[7-10].

In the modern drug discovery paradigm, which typically begins with the selection of a single molecular target, the ideal inhibitory drug is one that inhibits a single gene product so completely and so specifically that it is as if the gene product were absent. Treating cells with such a drug should induce changes in gene expression very similar to those resulting from deleting the gene encoding the drug's target. Here we have compared the genome-wide effects on gene expression that result from deletions of various genes in the budding yeast *S. cerevisiae* to the effects on gene expression that result from treatment

with known inhibitors of those gene products. Using the calcineurin signaling pathway as a model system, we tested an approach that permits identification of genes that encode proteins specifically involved in pathways affected by a drug. The FK506 characteristic pattern, or 'signature', of altered gene expression was not observed in mutant cells lacking proteins inhibited by FK506 (for example, a calcineurin or FK506-binding-protein mutant strain), but was observed in mutants deleted for genes in pathways unrelated to FK506 action (for example, a cyclophilin mutant strain). Conversely, the cyclosporin A (CsA) signature was not observed in CsA-treated calcineurin or cyclophilin mutant strains, but was seen in an FK506-binding-protein mutant strain treated with CsA. The method also demonstrates that FK506, a clinically used immunosuppressant, has 'off-target' effects that are independent of its binding to immunophilins. Thus, the approach we describe may provide a way to identify the pathways altered by a drug and to detect drug effects mediated through unintended targets.

Null mutants phenocopy drug-treated cells on a genomic scale
To test whether a null mutation in a drug target serves as a model of an ideal inhibitory drug, we examined the effects on gene expression associated with pharmacological or genetic inhibition of calcineurin function. Calcineurin is a highly conserved calcium- and calmodulin-activated serine/threonine protein phosphatase implicated in diverse processes dependent on calcium signaling[12-13]. In budding yeast, calcineurin is required for intracellular ion homeostasis[14], for adaptation to prolonged mating pheromone treatment[15] and in the regulation of

Fig. 1 Model of antagonism of the calcineurin signaling pathway mediated by FK506 and cyclosporin A (CsA). Calcineurin activity is composed of a catalytic subunit (calcineurin A, encoded in yeast by the CNA1 and CNA2 genes), and calcium-binding regulatory subunits calmodulin (CMD) and calcineurin B (CnB). After entering cells, FK506 and CsA specifically bind and inhibit the peptidyl-proline isomerase activity of their respective immunophilins, FK506 binding proteins (FKBP) and cyclophilins (CyP). The most abundant immunophilins in yeast (Fpr1 and Cph1) are thought to mediate calcineurin inhibition. Drug–immunophilin complexes bind and inhibit the calcium- and calmodulin-stimulated phosphatase calcineurin. Among the substrates of calcineurin are transcriptional activators that act to modulate gene expression.

the onset of mitosis[16]. In mammals, calcineurin has been implicated in T-cell activation[12], in apoptosis[17], in cardiac hypertrophy[18] and in the transition from short-term to long-term memory[19]. In both organisms, calcineurin activity is inhibited by FK506 and CsA, immunosuppressant drugs whose effects on calcineurin are mediated through families of intracellular receptor proteins called immunophilins[12,20] (Fig. 1). To assess the effects of pharmacologic inhibition of calcineurin, wild-type S. cerevisiae was grown to early logarithmic phase in the presence or absence of FK506 or CsA. Isogenic cells, from which the genes encoding the catalytic subunits of calcineurin (CNA1 and CNA2) had been deleted[21] (referred to as the cna or calcineurin mutant), were grown in parallel, in the absence of the drug. Fluorescently-labeled cDNA was prepared by reverse transcription of polyA⁺ RNA in the presence of Cy3- or Cy5-deoxynucleotide triphosphates and then hybridized to a microarray containing more than 6,000 DNA probes representing 97% of the known or predicted ORFs in the yeast genome. Simultaneous hybridization of Cy5-labeled cDNA from mock-treated cells and Cy3-labeled cDNA from cells treated with 1 µg/ml FK506 allowed the effect of drug treatment on mRNA levels of each ORF to be determined (Fig. 2a and b and data not shown). Similarly, effects of the calcineurin mutations on the mRNA levels of each gene were assessed by simultaneous hybridization of Cy5-labeled cDNA from wild-type cells and Cy3-labeled cDNA from the calcineurin mutant strain(Fig. 2c). For each comparison of this kind, reported expression ratios are the average of at least two hybridizations in which the Cy3 and Cy5 fluors were reversed to remove biases that may be introduced by gene-specific differences in incorporation of the two fluors (data not shown).

Treatment with FK506 in these growth conditions resulted in a signature pattern of altered gene expression in which mRNA levels of 36 ORFs changed by more than twofold (http://www.rosetta.org). A very similar pattern of altered gene expression was observed when the calcineurin mutant strain was compared to wild-type cells. Comparison of the changes in mRNA expression of each gene resulting from treatment of wild-type cells with FK506 with mRNA expression changes resulting from deletion of the calcineurin genes showed the considerable similarity of the global transcript alterations in response to the two perturbations (Fig. 2b–d). Quantification of this similarity using the correlation coefficient (ρ) showed large correlations between the FK506 treatment signature and the calcineurin deletion signature (ρ = 0.75 ± 0.03), as well as the CsA treatment signature (ρ = 0.94±0.02), but not with a randomly selected deletion mutant strain (deleted for the YER071C gene; ρ = -0.07 ± 0.04; Fig. 2e). The FK506 treatment signature was also compared with those of more than 40 other deletion mutant strains or drug-treatments thought to affect

unrelated pathways, and none had statistically significant correlations. These data establish that genetic disruption of calcineurin function provides a close and specific phenocopy of treatment with FK506 or CsA.

To avoid generalizing from a single example, we also compared the effects of treatment of wild-type cells with 3-aminotriazole (3-AT) with the effects of deletion of the HIS3 gene. HIS3 encodes imidazoleglycerol phosphate dehydratase, which catalyzes the seventh step of the histidine biosynthetic pathway in yeast[22]; 3-AT is a competitive inhibitor of this enzyme that triggers a large transcriptional amino-acid starvation response[23]. Microarray analysis of wild-type and isogenic his3-deficient strains demonstrated the expected large genome-wide transcriptional responses (involving more than 1,000 ORFs) resulting from treatment with 3-AT (Fig. 3a) or from HIS3 deletion (Fig. 3c). Quantitative comparison of the 3-AT treatment signature and the his3 mutant signature showed a high level of correlation (ρ= 0.76 ± 0.02) that even extended to genes that experienced small changes in expression level (Fig. 3b). As a negative control, the correlations between the 3-AT treatment signature or the his3 mutant signature and the calcineurin mutant strain were not statistically significant (ρ = 0.09 ± 0.06 and -0.01 ± 0.04, respectively). That both the calcineurin/FK506 and the his3/3-AT comparisons were highly correlated indicates that in many cases the expression profile resulting from a gene deletion closely resembles the expression profile of wild-type cells treated with an inhibitor of that gene's product.

'Decoder' strategy: Drug target validation with deletion mutants
Because pharmacological inhibition of different targets might give similar or identical expression profiles, simple comparison of drug signatures to mutant signatures is unlikely to unambiguously identify a drug's target. To overcome this limitation, an additional 'decoder' step is used. We first compare the expression profile of wild-type drug-treated cells to the expression profiles from a panel of genetic mutant strains, using a correlation coefficient metric. Mutant strains whose expression profile is similar to that of drug-treated wild-type cells are selected and subjected to drug treatment, generating the drug signature in the mutant strain (that is, the mutant drug signature). If the mutated gene encodes a protein involved in a pathway affected by the drug, we expect the drug signature in mutant cells to be different (or absent, for an ideal drug) from the drug signature seen in wild-type cells.

Fig. 2 Expression profiles from FK506-treated wild-type (wt) cells and a calcineurin-disruption mutant strain share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from FK506 treatment (*a* and *b*) or from genetic disruption of genes encoding calcineurin (*c*). *a*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated strain R563 and Cy3-labeled cDNA (green) from strain R563 treated with 1 µg/ml FK506. *b*, Enlarged view of the boxed area in *a*. Arrowheads indicate specific ORFs induced or repressed. *c*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from strain R563 and Cy3-labeled cDNA (green) from strain MCY300 (deleted for the *CNA1,CNA2* catalytic subunits of calcineurin). Arrows indicate specific ORFs induced or repressed. *d*, The log$_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the log$_{10}$ of the expression ratio in the calcineurin mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. *e*, The log$_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the log$_{10}$ of the expression ratio in the *yer071c* mutant hybridizations. No ORFs were induced or repressed in both experiments.

To illustrate this, we treated the *his3* mutant strain with 3-AT. The signature pattern of altered gene expression resulting from treatment of the mutant strain with 3-AT was much less complex than that of the 3-AT signature in wild-type cells (Fig. 4). This is seen simply by examining plots of mean intensity of the hybridization signal (which approximately reflects level of expression) versus the expression ratio for each ORF (Fig. 4). Genes that were expressed at higher or lower levels in 3-AT treated cells or in *his3* mutant cells are shown as red and green dots, respectively. We analyzed the 3-AT signature in wild-type (Fig. 4a) and *his3* mutant cells (Fig. 4c), as well as the *his3* mutant strain signature (Fig. 4b). Whereas histidine limitation induced by 3-AT induced more than 1,000 transcription-level changes in the wild-type strain, few or no transcript level changes were induced by treatment of the *his3*-deletion strain with 3-AT. This indicates that with the growth conditions used, essentially all of the effects of 3-AT depend on or are mediated through the HIS3 gene product.

Applying this approach to the calcineurin signaling pathway showed the specificity of the method. The calcineurin mutant strain and strains with deletions in the genes encoding the most abundant immunophilins in yeast[12] (*CPH1* and *FPR1*) were treated with either FK506 or CsA to determine the profiles of altered gene expression resulting from drug treatment of the mutant cells (that is, mutant +/- drug). We compared the drug signatures in the mutants to the wild-type drug signature using the correlation coefficient metric (Table 1). Although the signature generated by treatment of wild-type cells with FK506 was highly correlated to the calcineurin mutant strain signature (ρ = 0.75 ± 0.03), it bore no similarity to the profile after treatment of the calcineurin mutant strain with FK506 (ρ = -0.01 ± 0.07). This indicates that FK506 was unable to elicit its normal transcriptional response in the calcineurin mutant strain. Likewise, treatment of the *fpr1* mutant strain with FK506 elicited an expression profile that was not correlated to the FK506 signature in the wild-type strain (ρ = -0.23 ± 0.07), indicating that the *FPR1* gene product is likely to be involved in the pathway affected by FK506. The same was true for the *cna fpr1* mutant strain. In contrast, treatment of the *cph1* mutant strain with FK506 generated an expression profile highly correlated with the wild-type FK506 expression profile (ρ = 0.79 ± 0.03), indicating the *cph1* mutation did not block the mode of action of FK506 and thus is not directly involved in the pathway affected by FK506. We tabulated the change in expression in response to FK506 in different mutant strains for all ORFs with expression ratios greater than 1.8 in FK506-treated cells or in the calcineurin mutant strain (Fig. 5a).The calcineurin mutant strain signature and the FK506 responses in wild-type and the *cph1* mutant strain are similar, and there are no transcript-level changes (seen in black) for treatment of the calcineurin, *fpr1* and *cna fpr1* mutant strains with FK506 (Fig. 5a).

Similar experiments and analyses with CsA provided further validation of this approach. The expression profile elicited by treatment of wild-type cells with CsA was highly corre-

Table 1 Signature correlation of expression ratios as a result of FK506 treatment in various mutant strains

	wild-type +/-FK506	cna +/-FK506	fpr1 +/-FK506	cna fpr1 +/-FK506	cph1 +/-FK506
wild-type +/- FK506	0.93 ± 0.04	-0.01 ± 0.07	-0.23 ± 0.07	0.12 ± 0.07	0.79 ± 0.03

Signature correlation shows the absence of the FK506 signature specifically in the calcineurin (*cna*) and *fpr1* (major FK506 binding protein) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- FK506 experiments.

Fig. 3 Expression profiles from a *his3* mutant strain and wild-type (wt) cells treated with 3-AT share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from 3-AT treatment (*a*) or from genetic disruption of the *HIS3* gene (*c*). *a*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated wild-type strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. *b*, Plot of the $\log_{10}$ of the expression ratio for each ORF derived from the 3-AT treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the *his3* mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. The correlation of expression ratios applies not only to genes with large expression ratios (for example, *CHA1* and *ARG1*), but also extends to genes with expression ratios less than 2 (for example, *ILV1* and *CPH1*). *ILV1* is induced 1.9-fold and 1.5-fold, and *CPH1* is downregulated 1.9-fold

and 1.7-fold, in cells treated with 3-AT and *his3* mutant cells, respectively. Two ORFs do not fall on the line x = y. The leftmost point is the *HIS3* data point, which is induced by 3-AT treatment but which is not absent from the *his3* mutant strain. The other point is *YOR203w*. Both data points are labeled *HIS3* because hybridization to *YOR203w* is most likely due to *HIS3* mRNA, as *YOR203w* overlaps the *HIS3* open reading frame. *c*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from wild-type strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. Arrowheads indicate specific ORFs induced or repressed.

lated to the profile elicited by mutation of the calcineurin genes (ρ = 0.71 ± 0.04), but did not correlate with the expression profile resulting from treatment of the calcineurin mutant strain with CsA (ρ = -0.05 ± 0.07; Table 2), indicating that the genetic deletion of calcineurin interfered with the ability of CsA to elicit its normal transcriptional response. Likewise, the CsA signature was essentially absent in CsA-treated *cph1* mutant cells, and the expression profile of CsA-treated *cph1* mutant cells correlated poorly to that of CsA-treated wild-type cells (ρ = 0.18 ± 0.07). Thus, the *CPH1* gene product was required for the CsA response seen in wild-type cells. Conversely, treatment of *fpr1* mutant cells with CsA resulted in an expression pattern very similar to the profile of CsA-treated wild-type cells (ρ = 0.77 ± 0.03), indicating that *FPR1* was not necessary for the CsA-mediated effects. Analysis of individual ORFs affected by CsA and their expression ratios over the entire set of experiments confirmed that *CPH1* and the genes encoding calcineurin, but not

FPR1, are necessary for the wild-type CsA response (Fig. 5*b*). The observation that the profiles resulting from FK506 or CsA drug treatment are similar to that of the calcineurin deletion mutant strain might allow the prediction that calcineurin was involved in the pathway affected by these drugs. But because the expression profile of the *fpr1* mutant strain did not bear a strong similarity to the wild-type drug expression profile for FK506, it is obvious that the drug treatment of the mutant strains was necessary to identify Fpr1, but not Cph1, as a potential FK506 drug target. In the same way, the 'decoder' strategy was necessary to identify Cph1, but not Fpr1, as a potential drug target for CsA.

'Decoder' approach can identify secondary drug effects
For a drug that has a single biochemical target, the strategy outlined above may be useful in target validation. In many cases, however, a compound may affect multiple pathways and elicit a very complex signature. 'Decoding' such a complex signature

Fig. 4 Treatment of the *his3* mutant strain with 3-AT shows nearly complete loss of 3-AT signature. A plot of the $\log_{10}$ of the mean intensity of hybridization for each ORF versus the $\log_{10}$ of its expression ratio for each experiment is shown next to a pseudo-color image of a representative portion of the microarray. ORFs that are induced or repressed at the 95% confidence level are shown in green and red, respectively. *a*, Expression profile from treatment of the wild-type (wt) strain with 3-AT. Cy5-labeled cDNA (red) from mock-treated strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. *b*, Expression profile

from the *his3* deletion strain. Cy5-labeled cDNA (red) from strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. *c*, Expression profile of treatment of the *his3* deletion strain with 3-AT. Cy3-labeled cDNA (red) from *his3*-deleted strain R1226 and Cy5-labeled cDNA (green) from strain R1226 treated with 10 mM 3-AT. Arrowheads indicate the DNA probe and data point corresponding to the *HIS3* gene. The blue dashed line represents the threshold below which errors tend to increase rapidly because spot intensities are not sufficiently above background intensity.
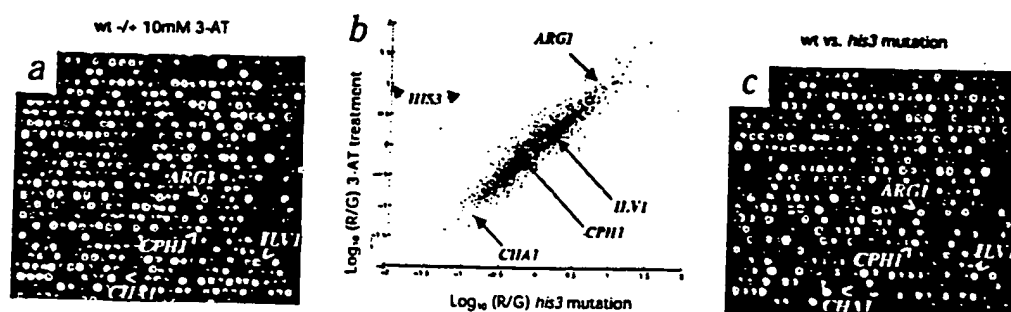
Table 2   Signature correlation of expression ratios as a result of CsA treatment in various mutant strains

	wild-type +/–CsA	cna +/–CsA	fpr1 +/–CsA	cna cph1 +/–CsA	cph1 +/–CsA
wild-type +/– CsA	0.94 ± 0.04	–0.05 ± .07	0.77 ± 0.03	–0.11 ± 0.07	0.18 ± 0.07

Signature correlation shows the absence of the CsA signature specifically in the calcineurin (cna) and cph1 (cyclophilin) deletion mutants. cna represents the mutant with deletions of the catalytic subunits of calcineurin, CNA1 and CNA2. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/– CsA experiments.

into the effects mediated through the intended target (the 'on-target signature') and those mediated through unintended targets (the 'off-target' signature) might be useful in evaluating a compound's specificity. Our 'decoder' strategy is based on the premise that 'off-target' signature should be insensitive to the genetic disruption of the primary target.

To determine whether the 'decoder' approach could identify an 'off-target' profile, we looked for a drug-responsive gene whose expression is insensitive to deletion of the primary target. To increase the likelihood of observing such genes, the same strains described in Tables 1 and 2 were treated with higher concentrations (50 μg/ml) of FK506. This led to a much more complex expression profile in wild-type cells, indicating that at this higher concentration, FK506 was inhibiting or activating additional targets. Several of the ORFs in this expanded FK506-induced expression profile were not affected by the calcineurin, cph1 or fpr1 mutations, as drug treatment of these mutant strains did not block their presence in the FK506 expression signature (Fig. 6). This indicates that FK506 was triggering changes in transcript levels of many genes through pathways independent of calcineurin, CPH1 and FPR1. Many of the upregulated ORFs in the 'off-target' pathway were genes reported to be regulated by the transcriptional activator Gcn4 (ref. 24). In some strains, a reporter gene under GCN4 control was induced in response to FK506 treatment[25]. To determine whether GCN4 is involved in this pathway that is independent of calcineurin, CPH1 and FPR1, we analyzed the effects of treatment with high-dose FK506 on global gene expression in a strain with a GCN4 deletion (Fig. 6). Of the 41 ORFs with calcineurin-independent expression ratios greater than 4, 32 were not induced in the gcn4 mutant, indicating that their induction by FK506 was GCN4-dependent. Not all GCN4-regulated genes were induced by FK506. This FK506-induced subset of GCN4-regulated genes may be those most sensitive to subtle changes in Gcn4 levels, or perhaps other regulatory circuits prevent FK506 activation of some GCN4-regulated genes. Seven of the remaining nine ORFs induced by FK506 were independent of

both the calcineurin and GCN4 pathways. The simplest explanation is that FK506 inhibits or activates additional pathways. Members of this class include SNQ2 and PDR5, genes that encode drug efflux pumps with structural homology to mammalian multiple drug resistance proteins[26]. FK506 may interact directly with Pdr5 to inhibit its function[27]. Our results indicate that treatment with FK506 leads to fourfold-to-sixfold induction of PDR5 mRNA levels. YOR1, another gene that can confer drug resistance, is also induced threefold-to-fourfold by FK506. Thus, drug treatment of strains with mutations in the primary targets can prove useful in identifying effects mediated by secondary drug targets, including the nature and extent of newly discovered and previously unsuspected pathways affected by the drug.

We describe here a method for drug target validation and the identification of secondary drug target effects that uses DNA microarrays to survey the effects of drugs on global gene expression patterns. We established that genetic and pharmacologic inhibition of gene function can result in extremely similar changes in gene expression. We also demonstrated that one can confirm a potential drug target by treating a deletion mutant defective in the gene encoding the putative target. Drug-mediated signatures from strains with mutations in pathways or processes directly or indirectly affected by the drug bore little or

Fig. 5   Response of FK506 and CsA signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 1.8 in response to treatment with 1 μg/ml FK506 (a) or 50 μg/ml CsA (b) are listed (left side) and their expression ratios in the indicated strain are shown on the green (induction)–red (repression) color scale. a, Calcineurin (cna) mutant and FK506 treatment signature genes are in the first two columns. Almost all FK506 signature genes have expression ratios near unity in deletion strains involved in pathways affected by FK506 (calcineurin, fpr1 and cna fpr1 mutants) but not in deletion strains in unrelated pathways (cph1). b, Calcineurin (cna) mutant and CsA treatment signature genes are in the first two columns. Almost all CsA signature genes have expression ratios near unity in deletion strains involved in pathways affected by CsA (calcineurin, cph1 and cna cph1 mutants) but not in deletion strains in unrelated pathways (fpr1).

Strain: wt cna fpr1 cph1 cna fpr1 gcn4

FK506: + + + + + +

Fold repression — Fold induction

Fig. 6 Response of FK506 signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 4 in at least one experiment are listed and their expression ratios in the indicated strain are shown in the green (induction)–red (repression) color scale. The genes have been divided into classes corresponding to these expected behaviors: 'CNA-dependent' genes respond to FK506 (50 µg/ml) except when either calcineurin genes or FPR1 or both are deleted; 'GCN4-dependent' genes respond to FK506 except when GCN4 is deleted. These genes still respond to FK506 when calcineurin genes or FPR1 or CPH1 are deleted; that is, their responses are not mediated by calcineurin, Cph1, or Fpr1. 'CNA- and GCN4-independent' genes respond to FK506 in all deletion strains tested. A 'complex behavior' class is provided for those genes that did not match the model of FK506 response mediated through calcineurin or Fpr1 or separately through Gcn4.

no similarity to the wild-type drug expression profile. In contrast, drug-mediated signatures from strains with mutations in genes involved in pathways unrelated to the drug's action showed extensive similarity to the wild-type drug signature. By applying this approach to a drug that affects multiple pathways (FK506), we were able to decode a complex signature into component parts, including the identification of an 'off-target' signature that was mediated through pathways independent of calcineurin or the Fpr1 immunophilin.

Discussion

It is well-established that high-throughput biochemical screening can identify potent inhibitory compounds against a given target. The 'decoder' approach described here complements this process by evaluating the equally important property of specificity: the tendency of a compound to inhibit pathways other than that of its intended target. The ability to observe such 'off-target' effects will likely be useful in several ways. Profiling compounds with known toxicities will allow the development of a database of expression changes associated with particular toxicities. Recognition of potential toxicities in the 'off-target' signatures of otherwise promising compounds then may allow earlier identification of those likely to fail in clinical trials. Comparing the extent and peculiarities of 'off-target' signatures of promising drug candiates could provide a new way to group compounds by their effects on secondary pathways, even before those effects are understood. This may prove to be an alternative, potentially more effective, way to select compounds for animal and clinical trials. Some drugs are more effective against a related protein than against the originally intended target. Sildenafil (Viagra™), for example, was initially developed as a phosphodiesterase inhibitor to control cardiac contractility, but was found to be highly specific for phosphodiesterase 5, an isozyme whose inhibition overcomes defects in

penile erection. It is possible that application of the 'decoder' to other compounds may show that they too have a potent activity against a target distinct from their intended target.

The ability to decode drug effects is dependent on the availability of functionally 'targetless' cells. In yeast, this is being achieved by systematically disrupting each yeast gene (Saccharomyces Deletion Consortium: http://sequence-www.stanford.edu/group/yeast_deletion_project/deletion.html). Efforts are underway to obtain expression profiles from each deletion mutant strain. Determining signatures resulting from inactivation of essential genes presents a unique problem, but it may be possible to do so by examining heterozygotes or by using a controllable promoter to reduce expression of the essential gene. Although it is already feasible to test several compounds in dozens of yeast strains, another challenge for the 'decoder' strategy will be the efficient selection of the mutants with deletions in genes most likely to encode the intended drug target. The signature correlation plots described are one metric that could be used as part of that selection process, but others need to be explored. Applying the 'decoder' to mammalian cells presents additional challenges. It is considerably more difficult to isolate functionally 'targetless' cells. Strategies involving titratable promoters, known specific inhibitors, anti-sense RNAs, ribozymes, and methods of targeting specific proteins for degradation are possible and should be tested. Another limitation is that not all cell types express the same set of genes and therefore 'off-target' effects may be different in different cell types. In addition, applying the 'decoder' to human cells will also require technical improvements that allow expression profiling from a small number of cells. Even the broader question of whether the insensitivity of 'off-target' signatures to the disruption of the main target is the exception or the rule can only be answered by the accumulation of more data. Barkai and Leibler, however, have argued in favor of robustness of biological networks, indicating that drug perturbations ('off-target' signatures) may be robust even when the system is subjected to another perturbation (such as a genetic disruption)(ref. 28). Many practical developments will be necessary if the 'decoder' concept is to be broadly applied.

Expression arrays have been used mainly as an initial screen for genes induced in a particular tissue or process of interest by focusing on genes with large expression ratios. We have found, however, that effort to refine experimental protocols and repeat experiments increases the reliability of the data and permits new applications. For example, it provides a larger set

1298

Table 3	Yeast strains used	
Strain	Relevant genotype	Reference
YPH499	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1*	(34)
R563	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3*	(this study)
R558	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 fpr1::HIS3*	(this study)
R567	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cph1::HIS3*	(this study)
MCY300	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3*	(21)
R132	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 cph1::kanr*	(this study)
R133	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 fpr1::kanr*	(this study)
R559	*Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3 gcn4::LEU2*	(this study)
BY4719	*Mata trp1-Δ63 ura3-Δ0*	(35)
BY4738	*Matα trp1-Δ 63 ura3-Δ0*	(35)
R491	*Mata/α BY4719 X BY4738*	(this study)
BY4728	*Mata his3-Δ200 trp1-Δ63 ura3-Δ0*	(35)
BY4729	*Matα his3-Δ 200 trp1-Δ63 ura3-Δ0*	(35)
R1226	*Mata/α BY4728 X BY4729*	(this study)

of genes at higher confidence levels that serve as a more unique signature for a given protein perturbation. In addition, it allows subtle signatures to be detected. when, for example, a protein is only partially inhibited. This may enable clinical monitoring of small changes in protein function in disease or toxicity states before they could otherwise be detected. Because the functions of many genes detected on transcript arrays are known, these microarrays are powerful tools that provide detailed information about a cell's physiology. For example, changes in the flux through a metabolic pathway are reflected in transcriptional changes in genes in the pathway[7]. Furthermore, it may be possible to indirectly measure protein activity levels from expression profiling data (S.F., *et al.*, unpublished data). Thus, although the eventual development of genomic methods allowing the direct measurement of all cellular protein levels will be an important achievement, transcript array technology offers an immediate and robust means of evaluating the effects of various treatments on gene expression and protein function.

Methods

Construction, growth and drug treatment of yeast strains. The strains used in this study (Table 3) were constructed by standard techniques[29]. To construct strain R559, strain R563 was transformed to Leu⁺ with plasmid pM12 digested by *SaI* and *MluI* (provided by A. Hinnebusch and T. Dever). Strains R132 and R133 were constructed by transforming the bacterial kanamycin resistance cassette[30] flanked by genomic DNA from the *CPH1* and *FPR1* loci, respectively, and selecting for G418-resistant colonies. For experiments with FK506, cells were grown for three generations to a density of 1 × 10⁷ cells/ml in YAPD medium (YPD plus 0.004% adenine) supplemented with 10 mM calcium chloride as described[31]. Where indicated, FK506 was added to a final concentration of 1 µg/ml 0.5 h after inoculation of the culture or to 50 µg/ml 1 h before cells were collected. CsA was used at a final concentration of 50 µg/ml. Cells were broken by standard procedures[32] with the following modifications: Cell pellets were resuspended in breaking buffer (0.2 M Tris HCl pH 7.6, 0.5 M NaCl, 10 mM EDTA, 1% SDS), vortexed for 2 min on a VWR multi-tube vortexer at setting 8 in the presence of 60% glass beads (425–600 µm mesh; Sigma) and phenol:chloroform (50:50, volume/volume). After separation of the phases, the aqueous phase was re-extracted and ethanol-precipitated. Poly A⁺ RNA was isolated by two sequential chromatographic purifications over oligo dT cellulose (New England Biolabs, Beverly, Massachusetts) using established protocols[33].

For experiments using 3-AT, wild-type or *his3/his3* cells were grown to early logarithmic phase in SC medium, pelleted and resuspended in SC medium lacking histidine for 1 hr in the presence or absence of 10 mM 3-

AT, as indicated. Cells were harvested and mRNA isolated as above. FK506 was obtained from the Swedish Hospital Pharmacy (Seattle, Washington) and purified to homogeneity by ethyl acetate extraction by J. Simon (Fred Hutchinson Cancer Research Center, Seattle, Washington). CsA was obtained from Alexis Biochemicals (San Diego, California); 3-AT was from Sigma.

Preparation and hybridization of the labeled sample. Fluorescently-labeled cDNA was prepared, purified and hybridized essentially as described[7]. Cy3- or Cy5-dUTP (Amersham) was incorporated into cDNA during reverse transcription (Superscript II; Life Technologies) and purified by concentrating to less than 10 µl using Microcon-30 microconcentrators (Amicon, Houston, Texas). Paired cDNAs were resuspended in 20–26 µl hybridization solution (3 × SSC, 0.75 µg/ml polyA DNA, 0.2% SDS) and applied to the microarray under a 22- × 30-mm coverslip for 6 h at 63 °C, all according to a published method[7].

Fabrication and scanning of microarrays. PCR products containing common 5′ and 3′ sequences (Research Genetics, Huntsville, Alabama) were used as templates with amino-modified forward primer and unmodified reverse primers to PCR amplify 6,065 ORFs from the *S. cerevisiae* genome. Our first-pass success rate was 94%. Amplification reactions that gave products of unexpected sizes were excluded from subsequent analysis. ORFs that could not be amplified from purchased templates were amplified from genomic DNA. DNA samples from 100-µl reactions were isopropanol-precipitated, resuspended in water, brought to a final concentration of 3× SSC in a total volume of 15 µl, and transferred to 384-well microtiter plates (Genetix Limited, Christchurch, Dorset, England). PCR products were spotted onto 1 × 3-inch polylysine-treated glass slides by a robot built essentially according to defined specifications[2,6,7] (http://cmgm.stanford.edu/pbrown/MGuide). After being printed, slides were processed according to published protocols[7].

Microarrays were imaged on a prototype multi-frame CCD camera in development at Applied Precision (Issaquah, Washington). Each CCD image frame was approximately 2-mm square. Exposure times of 2 s in the Cy5 channel (white light through Chroma 618–648 nm excitation filter, Chroma 657–727 nm emission filter) and 1 s in the Cy3 channel (Chroma 535–560 nm excitation filter, Chroma 570–620 nm emission filter) were done consecutively in each frame before moving to the next, spatially contiguous frame. Color isolation between the Cy3 and Cy5 channels was about 100:1 or better. Frames were 'knitted' together in software to make the complete images. The intensity of spots (about 100 µm) were quantified from the 10-µm pixels by frame-by-frame background subtraction and intensity averaging in each channel. Dynamic range of the resulting spot intensities was typically a ratio of 1,000 between the brightest spots and the background-subtracted additive error level. Normalization between the channels was accomplished by normalizing each channel to the mean intensities of all genes. This procedure is nearly equivalent to normalization between channels using the intensity

ratio of genomic DNA spots', but is possibly more robust, as it is based on the intensities of several thousand spots distributed over the array.

Signature correlation coefficients and their confidence limits. Correlation coefficients between the signature ORFs of various experiments were calculated using:

$$\rho = \sum_k x_k y_k \, / \, (\sum_k x_k^2 \sum_k y_k^2)^{1/2}$$

where x_k is the $\log_{10}$ of the expression ratio for the kth gene in the x signature, and y_k is the $\log_{10}$ of the expression ratio for the kth gene in the y signature. The summation is over those genes that were either up- or down-regulated in either experiment at the 95% confidence level. These genes each had a less than 5% chance of being actually unregulated (having expression ratios departing from unity due to measurement errors alone). This confidence level was assigned based on an error model which assigns a lognormal probability distribution to each gene's expression ratio with characteristic width based on the observed scatter in its repeated measurements (repeated arrays at the same nominal experimental conditions) and on the individual array hybridization quality. This latter dependence was derived from control experiments in which both Cy3 and Cy5 samples were derived from the same RNA sample. For large numbers of repeated measurements the error reduces to the observed scatter. For a single measurement the error is based on the array quality and the spot intensity.

Random measurement errors in the x and y signatures tend to bias the correlation towards zero. In most experiments, most genes are not significantly affected but do show small random measurement errors. Selecting only the '95% confidence' genes for the correlation calculation, rather than the entire genome, reduces this bias and makes the actual biological correlations more apparent.

Correlations between a profile and itself are unity by definition. Error limits on the correlation are 95% confidence limits based on the individual measurement error bars, and assuming uncorrelated errors[33]. They do not include the bias mentioned above; thus, a departure of ρ from unity does not necessarily mean that the underlying biological correlation is imperfect. However, a correlation of 0.7 ± 0.1, for example, is very significantly different from zero. Small (magnitude of ρ < 0.2) but formally significant correlation in the tables and text probably are due to small systematic biases in the Cy5/Cy3 ratios that violate the assumption of independent measurement errors used to generate the 95% confidence limits. Therefore, these small correlation values should be treated as not significant. A likely source of uncorrected systematic bias is the partially corrected scanner detector nonlinearity that differently affects the Cy3 and Cy5 detection channels.

The 1 μg/ml FK506 treatment signature was compared with more than 40 unrelated deletion mutant strain or drug signatures. These control profiles had correlation coefficients with the FK506 profile that were distributed around zero (mean ρ = –0.03) with a standard deviation of 0.16 (data not shown), and none had correlations greater than ρ = 0.38. Similarly, the calcineurin mutant strain signature correlated well with the CsA treatment signature (ρ = 0.71 ± 0.04) but not with the signatures from the negative controls (mean ρ = –0.02 with a standard deviation of 0.18).

Quality controls. End-to-end checks on expression ratio measurement accuracy were provided by analyzing the variance in repeated hybridizations using the same mRNA labeled with both Cy3 and Cy5, and also using Cy3 and Cy5 mRNA samples isolated from independent cultures of the same nominal strain and conditions. Biases undetected with this procedure, such as gene-specific biases presumably due to differential incorporation of Cy3- and Cy5-dUTP into cDNA, were minimized by doing hybridizations in fluor-reversed pairs, in which the Cy3/Cy5 labeling of the biological conditions was reversed in one experiment with respect to the other. The expression ratio for each gene is then the ratio of ratios between the two experiments in the pair. Other biases are removed by algorithmic numerical de-trending. The magnitude of these biases in the absence of de-trending and fluor reversal is typically about 30% in the ratio, but may be as high as twofold for some ORFs.

Expression ratios are based on mean intensities over each spot. Some smaller spots have fewer image pixels in the average. This does not degrade accuracy noticeably until the number of pixels falls below ten. In which case the spot is rejected from the data set. 'Wander' of spot positions with respect to the nominal grid is adaptively tracked in array subregions by the image processing software. Unequal spot 'wander' within a subregion greater than half-a-spot spacing is a difficulty for the automated quantitating algorithms; in this case, the spot is rejected from analysis based on human inspection of the 'wander'. Any spots partially overlapping are excluded from the data set. Less than 1% of spots typically are rejected for these reasons.

1. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470 (1995).
2. Schena, M. et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93, 10614–10619 (1996).
3. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645 (1996).
4. Lockhart, D.J. et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14, 1675–1680 (1996).
5. DeRisi, J. et al. Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* 14, 457–460 (1996).
6. Heller, R.A. et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94, 2150–2155 (1997).
7. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686 (1997).
8. Lashkari, D.A. et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94, 13057–13062 (1997).
9. Wodicka, L., Dong H., Mittman, M. Ho, M.-H. & Lockhart, D.J.. Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnol.* 15, 1359–1367 (1997).
10. Cho, R.J. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73 (1998).
11. Gray, N.S. et al. Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533–538 (1998).
12. Cardenas, M.E., Lorenz, M., Hemenway, C. & Heitman, J. Yeast as model T cells. *Perspect. Drug Discovery Design* 2, 103–126 (1994).
13. Klee, C.B., Ren, H. & Wang, X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. *J. Biol. Chem.* 273, 13367–13370 (1998).
14. Tanida, I., Hasegawa, A., Iida, H., Ohya, Y. & Anraku, Y. Cooperation of calcineurin and vacuolar H(+)-ATPase in intracellular Ca2+ homeostasis of yeast cells. *J. Biol. Chem.* 270, 10113–10119 (1995).
15. Moser, M.J., Geiser, J.R. & Davis, T.N. Ca2+-calmodulin promotes survival of pheromone-induced growth arrest by activation of calcineurin and Ca2+-calmodulin-dependent protein kinase. *Mol. Cell. Biol.* 16, 4824–4831 (1996).
16. Mizunuma, M., Hirata, D., Miyahara, K., Tsuchiya, E. & Miyakawa, T. Role of calcineurin and Mpk1 in regulating the onset of mitosis in budding yeast. *Nature* 392, 303–306 (1998).
17. Yazdanbakhsh, K., Choi, J.W., Li, Y., Lau, L.F. & Choi, Y. Cyclosporin A blocks apoptosis by inhibiting the DNA binding activity of the transcription factor Nur77. *Proc. Natl. Acad. Sci. USA* 92, 437–441 (1995).
18. Molkentin, J.D. et al. A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* 93, 215–228 (1998)
19. Mansuy, I.M., Mayford, M., Jacob, B., Kandel, E.R. & Bach, M.E. Restricted and regulated overexpression reveals calcineurin as a key component in the transition from short-term to long-term memory. *Cell* 92, 39–49 (1998).
20. Schreiber, S.L. & Crabtree, G.R. The mechanism of action of cyclosporin A and FK506. *Immunol. Today* 13, 136–142 (1992).
21. Cyert, M.S., Kunisawa, R., Kaim, D. & Thorner, J. Yeast has homologs (CNA1 and CNA2 gene products) of mammalian calcineurin, a calmodulin-regulated phosphoprotein phosphatase. *Proc. Natl. Acad. Sci. USA* 88, 7376–7380 (1991).
22. Jones, E.W. & Fink, G.R. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds. Strathern, J.N., Jones, E.W. & Broach, J.R.) 181–299 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
23. Hinnebusch, A. Translational regulation of yeast GCN4. *J. Biol. Chem.* 272, 21661–21664 (1997).
24. Hinnebusch, A.G. in *The Molecular and Cellular Biology of the Yeast*

Saccharomyces: Gene Expression. (eds. Jones, E.W., Pringle, J.R. & Broach, J.R.) 319–414 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).

25. Heitman, J. *et al.* The immunosuppressant FK506 inhibits amino acid import in *Saccharomyces cerevisiae. Mol. Cell. Biol.* 13, 5010–5019 (1993).
26. Balzi, E. & Goffeau, A. Yeast multidrug resistance: the PDR network. *J. Bioenerg. Biomembr.* 27, 71–76 (1995).
27. Egner, R., Rosenthal, F.E., Kralli, A., Sanglard, D. & Kuchler, K. Genetic separation of FK506 susceptibility and drug transport in the yeast Pdr5 ATP-binding cassette multidrug resistance transporter *Mol. Biol. Cell* 9, 523–543 (1998).
28. Barkai, N. & Leiber, S. Robustness in simple biochemical networks. *Nature* 387, 913–917 (1997).
29. Schiestl, R.H., Manivasakam, P., Woods, R.A. & Gietz, R.D. Introducing DNA into yeast by transformation. *Methods: A companion to Methods in Enzymology* 5, 79–85 (1993).
30. Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. New heterologous mod-

ules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae. Yeast* 10, 1793–1808 (1994).
31. Garrett-Engele, P., Moilanen, B. & Cyert, M.S. Calcineurin, the Ca2+/calmodulin-dependent protein phosphatase, is essential in yeast mutants with cell integrity defects and in mutants that lack a functional vacuolar H(+)-ATPase. *Mol. Cell. Biol.* 15, 4103–4114 (1995).
32. Ausubel, F.M. *et al.* in *Current Protocols in Molecular Biology* 13.12.1–13.12.5 (eds. Ausubel, F.M., *et al.*)(John Wiley & Sons, New York, 1993).
33. Bulmer, M.G. in *Principles of Statistics* 224–225 (Dover Publications, New York, 1979).
34. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains designated for efficient manipulation of DNA in *Saccharomyces cerevisiae. Genetics* 122, 19–27 (1989).
35. Brachmann, C.B. *et al.* Designer deletion strains derived from *Saccharomyces cerevisiae* S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115–132 (1998).

REPORTS

co mosaic viral RNA was obtained by phenol and chloroform extractions of the virus and precipitated from ethanol. CA-NC assembly reactions in the presence of noncognate RNAs were identical to those given in (9). In the absence of RNA, CA-NC cones formed under the following conditions: 300 μM CA-NC, 1 M NaCl, and 50 mM tris-HCl (pH 8.0) at 37°C for 60 min. In the absence of exogenous RNA, neither cones nor cylinders formed at concentrations of 0.5 M NaCl or below. Absorption spectra demonstrated that our CA-NC preparations were not contaminated with *Escherichia coli* RNA (estimated lower detection limit was ~1 base/protein molecule). To control for even lower levels of RNA contamination, we preincubated the CA-NC protein with 0.5 mg/ml ribonuclease A (Type 1-AS, 54 Kunitz U/mg, Sigma) for 1 hour at 4°C, which then formed cones normally.

13. V. Y. Klishko, data not shown.
14. M. Ge and K. Sattler, *Chem. Phys. Lett.* 220, 192 (1994).
15. A. Krishnan et al., *Nature* 388, 451 (1997).
16. L. B. Kong et al., *J. Virol.* 72, 4403 (1998).

17. Assembly mixtures were deposited on holey carbon grids, blotted briefly with filter paper, plunged into liquid ethane, and transferred to liquid nitrogen. Frozen grids were transferred to a Philips 420 TEM equipped with a Gatan cold stage system, and images of particles in vitreous ice were recorded under low dose conditions at 36,000× magnification and ~1.6-μm defocus.
18. J. T. Finch, data not shown.
19. R. A. Crowther, *Proceedings of the Third John Innes Symposium* (1976), pp. 15–25; E. Kellenberger, M. Häner, M. Wurtz, *Ultramicroscopy* 9, 139 (1982); J. Seymore and D. J. DeRosier, *J. Microsc.* 148, 195 (1987).
20. M. V. Nermut, C. Grief, S. Hashmi, D. J. Hockley, *AIDS Res. Hum. Retroviruses* 9, 929 (1993); M. V. Nermut et al., *Virology* 198, 288 (1994); E. Barklis, J. McDermott, S. Wilkens, S. Fuller, D. Thompson, *J. Biol. Chem.* 273, 7177 (1998); E. Barklis et al., *EMBO J.* 16, 1199 (1997); M. Yeager, E. M. Wilson-Kubalek, S. C. Weiner, P. O. Brown, A. Rein, *Proc. Natl. Acad. Sci. U.S.A.* 95, 7299 (1998).

21. J. T. Finch et al., unpublished observations.
22. V. M. Vogt, in (2), pp. 27–70.
23. M. A. McClure, M. S. Johnson, D.-F. Feng, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* 85, 2469-2473 (1988).
24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
25. We thank C. Hill for very helpful discussions on the relationship between viral cores and fullerene cones, D. Hobbs for refining the ChemDraw3D images of cones, G. Stubbs for a gift of tobacco mosiac virus, J. McCutcheon for the plasmid used to prepare ribosomal RNA, and K. Albertine and N. Chandler of the University of Utah Shared Electron Microscopy facility for their support and encouragement. Supported by grants from NIH and from the Huntsman Cancer Institute (to W.I.S.).

29 September 1998; accepted 17 November 1998

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

The temporal program of gene expression during a model physiological response of human cells, the response of fibroblasts to serum, was explored with a complementary DNA microarray representing about 8600 different human genes. Genes could be clustered into groups on the basis of their temporal patterns of expression in this program. Many features of the transcriptional program appeared to be related to the physiology of wound repair, suggesting that fibroblasts play a larger and richer role in this complex multicellular response than had previously been appreciated.
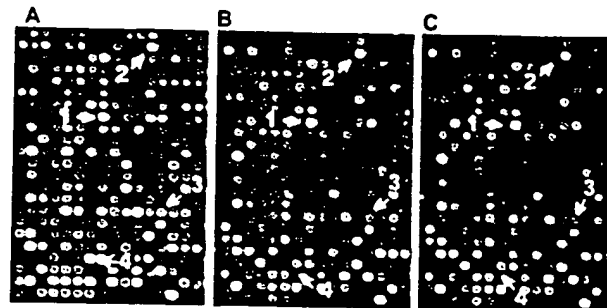
The response of mammalian fibroblasts to serum has been used as a model for studying growth control and cell cycle progression (1). Normal human fibroblasts require growth factors for proliferation in culture; these growth factors are usually provided by fetal bovine serum (FBS). In the absence of growth factors, fibroblasts enter a nondividing state, termed G_0, characterized by low metabolic activity. Addition of FBS or purified growth factors induces proliferation of the fibroblasts; the changes in gene expression that accompany this proliferative response have been the subject of many studies, and the responses of dozens of genes to serum have been characterized.

We took a fresh look at the response of human fibroblasts to serum, using cDNA microarrays representing about 8600 distinct human genes to observe the temporal program of transcription that underlies this response. Primary cultured fibroblasts from human neonatal foreskin were induced to enter a quiescent state by serum deprivation for 48 hours and then stimulated by addition of medium containing 10% FBS (2). DNA microarray hybridization was used to measure the temporal changes in mRNA levels of 8613 human genes (3) at 12 times, ranging from 15 min to 24 hours after serum stimulation. The cDNA made from purified mRNA from each sample was labeled with the fluorescent dye Cy5 and mixed with a common reference probe consisting of cDNA made from purified mRNA from the quiescent

V. R. Iyer and D. T. Ross, Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA. M. B. Eisen and D. Botstein, Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA. G. Schuler and M. S. Boguski, National Center for Biotechnology Information, Bethesda MD 20894, USA. T. Moore and J. Hudson Jr., Research Genetics, Huntsville, AL 35801, USA. J. C. F. Lee, D. Lashkari, D. Shalon, Incyte Pharmaceuticals, Fremont, CA 94555, USA. J. M. Trent, Laboratory of Cancer Genetics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. L. M. Staudt, Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, Bethesda, MD 20892, USA. P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford CA 94305, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

Fig. 1. The same section of the microarray is shown for three independent hybridizations comparing RNA isolated at the 8-hour time point after serum treatment to RNA from serum-deprived cells. Each microarray contained 9996 elements, including 9804 human cDNAs, representing 8613 different genes. mRNA from serum-deprived cells was used to prepare cDNA labeled with Cy3-deoxyuridine triphosphate (dUTP), and mRNA harvested from cells at different times after serum stimulation was used to prepare cDNA labeled with Cy5-dUTP. The two cDNA probes were mixed and simultaneously hybridized to the microarray. The image of the subsequent scan shows genes whose mRNAs are more abundant in the serum-deprived fibroblasts (that is, suppressed by serum treatment) as green spots and genes whose mRNAs are more abundant in the serum-treated fibroblasts as red spots. Yellow spots represent genes whose expression does not vary substantially between the two samples. The arrows indicate the spots representing the following genes: 1, protein disulfide isomerase-related protein P5; 2, IL-8 precursor; 3, EST AA057170; and 4, vascular endothelial growth factor.

culture (time zero) labeled with a second fluorescent dye, Cy3 (4). The color images of the hybridization results (Fig. 1) were made by representing the Cy3 fluorescent image as green and the Cy5 fluorescent image as red and merging the two color images.
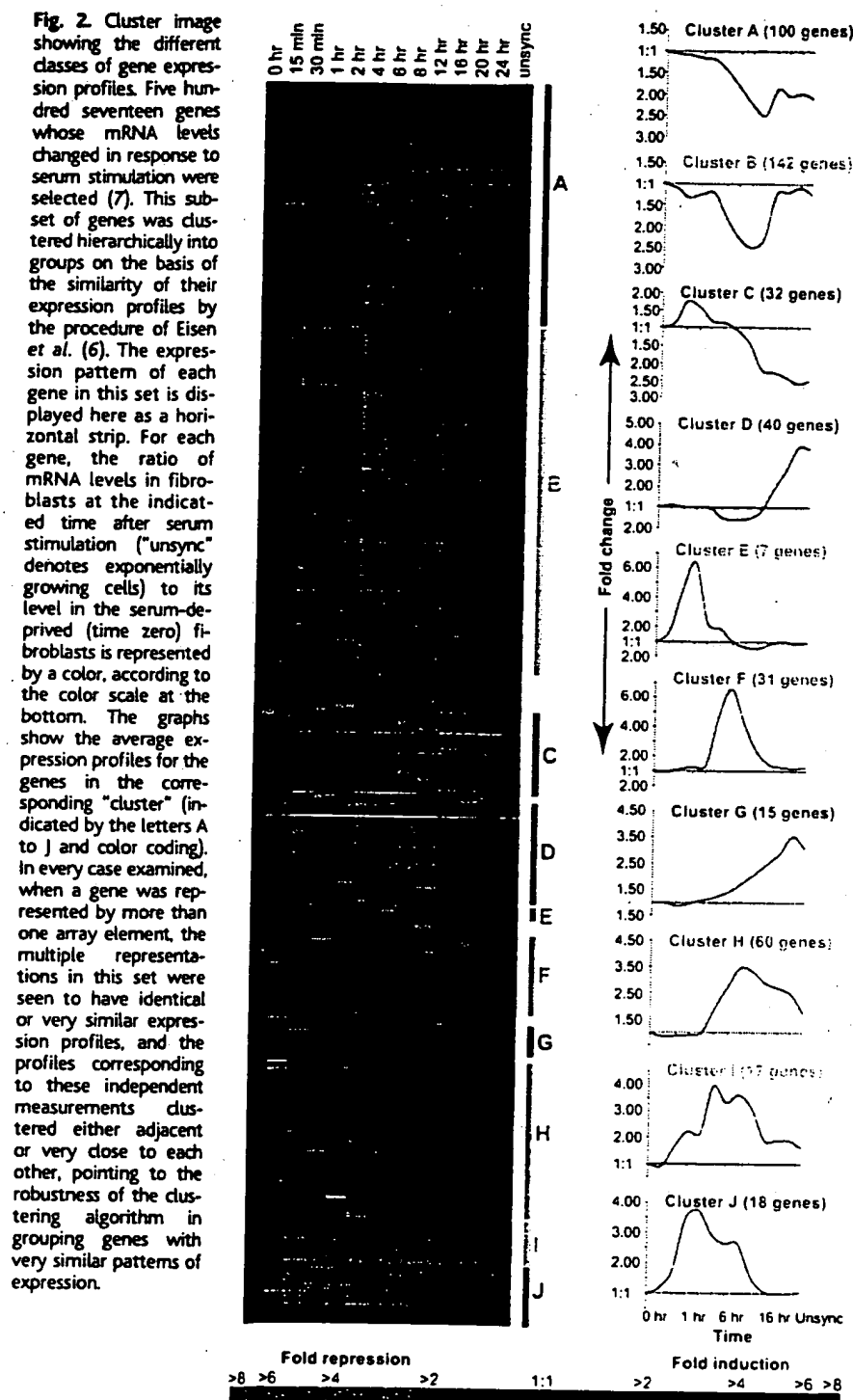
Diverse temporal profiles of gene expression could be seen among the 8613 genes surveyed in this experiment (Fig. 2); many of these genes (about half) were unnamed expressed sequence tags (ESTs) (5). Although diverse patterns of expression were observed, the orderly choreography of the expression program became apparent when the results were analyzed by a clustering and display method developed in our laboratory for analyzing genome-wide gene expression data (6). An example of such an analysis, here applied to a subset of 517 genes whose expression changed substantially in response to serum (7), is sh wn in Fig. 2. The entire detailed data set underlying Fig. 2 is available as a tab-delimited table (in cluster order) at the Science Web site (www. sciencemag.org/feature/data/984559.shl). In addition, the entire, larger data set for the complete set of genes analyzed in this experiment can be found at a Web site maintained by our laboratory (genome-www.stanford. edu/serum) (8).

One measure of the reliability of the changes we observed is inherent in the expression profiles of the genes. For most genes whose expression levels changed, we could see a gradual change over a few time points, which thus effectively provided independent measurements for almost all of the observations. An additional check was provided by the inclusion of duplicate and, in a few cases, multiple array elements representing the same gene for about 5% of the genes included in this microarray. In addition, three independent hybridizations to different microarrays with mRNA samples from cells harvested 8 hours after serum addition showed good correlation (Fig. 1). As an independent test, we measured the expression levels of several genes using the TaqMan 5' nuclease fluorigenic quantitative polymerase chain reaction (PCR) assay (9). The expression profiles of the genes, as measured by these two independent methods, were very similar (Fig. 3) (10).

The transcriptional response of fibroblasts to serum was extremely rapid. The immediate response to serum stimulation was dominated by genes that encode transcription factors and other proteins involved in signal transduction. The mRNAs for several genes [including c-FOS, JUN B, and mitogen-activated protein (MAP) kinase phosphatase-1 (MKP1)] were detectably induced within 15 min after serum stimulation (Fig. 4, A and B). Fifteen of the genes that were observed to be induced by serum encode known or suspected regulators of transcription (Fig. 4B). All but one were immediate-early genes—their induction was not inhibited by cycloheximide (11). This class of genes could be distinguished into those whose induction was transient (Fig. 2, cluster E) and those whose mRNA levels remained induced for much longer (Fig. 2, clusters I and J). Some features of the immediate response appeared to be directed at adaptation to the initiating signals. We observed a marked induction of mRNA encoding MKP1, a dual-specificity phosphatase that modulates the activity of the ERK1 and ERK2 MAP kinases (12). The coincidence of the peak of expression of genes in cluster E (Fig. 2) with that of MKP1 (Fig. 4A) suggests the possibility

Fig. 2. Cluster image showing the different classes of gene expression profiles. Five hundred seventeen genes whose mRNA levels changed in response to serum stimulation were selected (7). This subset of genes was clustered hierarchically into groups on the basis of the similarity of their expression profiles by the procedure of Eisen et al. (6). The expression pattern of each gene in this set is displayed here as a horizontal strip. For each gene, the ratio of mRNA levels in fibroblasts at the indicated time after serum stimulation ("unsync" denotes exponentially growing cells) to its level in the serum-deprived (time zero) fibroblasts is represented by a color, according to the color scale at the bottom. The graphs show the average expression profiles for the genes in the corresponding "cluster" (indicated by the letters A to J and color coding). In every case examined, when a gene was represented by more than one array element, the multiple representations in this set were seen to have identical or very similar expression profiles, and the profiles corresponding to these independent measurements clustered either adjacent or very close to each other, pointing to the robustness of the clustering algorithm in grouping genes with very similar patterns of expression.

that continued activity of the MAP kinase pathway is required to maintain induction of these genes but not of those with sustained expression (clusters I and J). The gene encoding a second member of the dual-specificity MAP kinase phosphatase family, known as dual-specificity protein phosphatase 6/pyst2, was induced later, at about 4 hours after serum stimulation. Genes encoding diverse other proteins with roles in signal transduction, ranging from cell-surface receptors [for example, the sphingosine 1-phosphate receptor (EDG-1), the vascular endothelial growth factor receptor, and the type II BMP receptor] to regulators of G-protein signaling (for example, NET1/p115 rho GEF) to DNA-binding transcription factors, were induced by serum (Fig. 4A).

The reprogramming of the regulatory circuits in response to serum involved not only induction of transcription factors but also reduced expression of many transcriptional regulators—some of which may play roles in maintaining the cells in G_0 or in priming them to react to wounding (Fig. 4C). Perhaps as a consequence of the historical focus on genes induced by serum stimulation of fibroblasts, the set of transcription factors whose expression diminished upon serum stimulation has been less well characterized.

Genes known or likely to be involved in controlling and mediating the proliferative response showed distinctive patterns of regulation. Several genes whose products inhibit progression of the cell-division cycle, such as p27 Kip1, p57 Kip2, and p18, were expressed in the quiescent fibroblasts and down-regulated before the onset of cell division. The nadir in the mRNA levels for these genes occurred between 6 and 12 hours after serum stimulation (Fig. 5A), coincident with the passage of the fibroblasts through G_1. The levels of the transcript encoding the WEE1-like protein kinase, which is believed to inhibit mitosis by phosphorylation of Cdc2, diminished between 4 and 8 to 12 hours after serum addition (Fig. 5A), well

before the onset of M phase at around 16 hours, raising the possibility of an additional role for Wee1 in an earlier stage of the cell cycle or in regulating the G_0 to G_1 transition. Several genes induced in the first few hours after serum stimulation, such as the helix-loop-helix proteins ID2 and ID3 and EST AA016305, a gene with homology to G_1-S cyclins, are candidates for roles in promoting the exit from G_0.

Genes involved in mediating progression through the cell cycle were characterized by a distinctive pattern of expression (Fig. 2, cluster D), reflecting the coincidence of their expression with the reentry of the stimulated fibroblasts into the cell-division cycle. The stimulated fibroblasts replicated their DNA about 16 hours after serum treatment. This timing was reflected by the induction of mRNA encoding both subunits of ribonucleotide reductase and PCNA, the processivity factor for DNA polymerase epsilon and delta. Cyclin A, Cyclin B1, Cdc2, and CDC28 kinase, regulators of passage through the S phase and the transition from G_2 to M phase, were induced at about 16 to 20 hours after serum addition. The kinase in the Cyclin B1–CDK pair needs to be activated by phosphorylation. The gene encoding Cyclin-dependent kinase 7 (CDK7: a homolog of *Xenopus* MO15 cdk-activating kinase) was induced in parallel with the Cdc2 and Cdc28 kinases (Fig. 5A), suggesting a potential role for CDK7 in mediating M phase. DNA topoisomerase II α, required for chromosome segregation at mitosis; Mad2, a component of the spindle checkpoint that prevents completion of mitosis (anaphase) if chromosomes are not attached to the spindle; and the kinetochore protein CENP-F all showed a similar expression profile.

In the hours after the serum stimulus, one of the most striking features of the unfolding transcriptional program was the appearance of numerous genes with known roles in processes relevant to the physiology of wound healing.

These included both genes involved in the direct role played by fibroblasts in remodeling of the clot and the extracellular matrix and, more notably, genes encoding proteins involved in intercellular signaling (Fig. 5). Genes induced in this program encode products that can (i) participate in the dynamic process of clotting, clot dissolution, and remodeling and perhaps contribute to hemostasis by promoting local vasoconstriction (for example, endothelin-1); (ii) promote chemotaxis and activation of neutrophils (for example, COX2) and recruitment and extravasation of monocytes and macrophages (for example, MCP1); (iii) promote chemotaxis and activation of T lymphocytes [for example, interleukin-8 (IL-8)] and B lymphocytes (for example, ICAM-1), thus providing both innate and antigen-specific defenses against wound infection and recruiting the phagocytic cells that will be required to clear out the debris during remodeling of the wound; (iv) promote angiogenesis and neovascularization (for example, VEGF) through newly forming tissue; (v) promote migration and proliferation of fibroblasts (for example, CTGF) and their differentiation into myofibroblasts (for example, Vimentin); and (vi) promote migration and proliferation of keratinocytes, leading to reepithelialization of the wound (for example, FGF7), and promote proliferation of melanocytes, perhaps contributing to wound hyperpigmentation (for example, FGF2).

Coordinated regulation of groups of genes whose products act at different steps in a common process was a recurring theme. For example, Furin, a prohormone-processing protease required for one of the processing steps in the generation of active endothelin, was induced in parallel with induction of the gene encoding the precursor of endothelin-1 (Fig. 5E) (13). Conversely, expression of CALLA/CD10, a membrane metalloprotease that degrades endothelin-1 and other peptide mediators of acute inflammation, was re-

Fig. 3. Independent verification of microarray quantitation. Relative mRNA levels of the indicated genes (Mast, mast/stem cell growth factor receptor) were measured with the TaqMan 5' nuclease fluorigenic quantitative PCR assay (9) (left) in the same samples that were used to prepare probes for microarray hybridizations (right). Data from the TaqMan analysis were normalized to mRNA concentrations and plotted relative to the level at time zero, so that the results could be compared with those from the microarray hybridizations. In general, quantitation with the two methods gave very similar results (10).

duced. A second example is provided by a set of five genes involved in the biosynthesis of cholesterol (Fig. 5I). The mRNAs encoding each of these enzymes showed sharply diminished expression beginning 4 to 6 hours after serum stimulation of fibroblasts. A likely explanation for the coordinated down-regulation of the cholesterol biosynthetic pathway is that serum provides cholesterol to fibroblasts through low-density lipoproteins, whereas in the absence of the cholesterol provided by serum, endogenous cholesterol biosynthesis in fibroblasts is required.

Many of the previously studied genes that we observed to be regulated in this program have no recognized role in any aspect of wound healing or fibroblast proliferation. Their identification in this study may therefore point to previously unknown aspects of these processes. A few selected genes in this group are shown in Fig. 5H. The stanniocalcin gene, for example (Fig. 5H), encodes a secreted protein without a clearly identified function in human cells (14, 15). Its induction in serum-stimulated fibro-

blasts suggests the possibility that it may play a role in the wound-healing process, perhaps serving as a signal in mediating inflammation or angiogenesis.

One of the most important results of this exploration was the discovery of over 200 previously unknown genes whose expression was regulated in specific temporal patterns during the response of fibroblasts to serum. For example, 13 of the 40 genes in cluster D (Fig. 2) have descriptive names that reflect their putative function. Nine of these 13 genes (69%) encode proteins that play roles in cell cycle progression, particularly in DNA replication and the G_2-M transition. This enrichment for cell cycle–related genes suggests that some of the

unnamed genes in this cluster—for example, EST W79311 and EST R13146, neither of which have sequence similarity to previously characterized genes—may represent previously unknown genes involved in this part of the cell cycle. Similarly, a remarkable fraction of genes that were grouped into cluster F on the basis of their expression profiles encoded proteins involved in intercellular signaling (Fig. 2), suggesting that a similar role should be considered for the many unnamed genes in this cluster. A disproportionately large fraction of the genes whose transcription diminished upon serum stimulation were unnamed ESTs.

Our intention was to use this experiment as a model to study the control of the transition

Fig. 4. "Reprogramming" of fibroblasts. Expression profiles of genes whose function is likely to play a role in the reprogramming phase of the response are shown with the same representation as in Fig. 2. In the cases in which a gene was represented by more than one element in the microarray, all measurements are shown. The genes were grouped into categories on the basis of our knowledge of their most likely role. Some genes with pleiotropic roles were included in more than one category.

Fig. 5. The transcriptional response to serum suggests a multifaceted role for fibroblasts in the physiology of wound healing. The features of the transcriptional program of fibroblasts in response to serum stimulation that appear to be related to various aspects of the wound-healing process and fibroblast proliferation are shown with the same convention for representing changes in transcript levels as was used in Figs. 2 and 4. (A) Cell cycle and proliferation, (B) coagulation and hemostasis, (C) inflammation, (D) angiogenesis, (E) tissue remodeling, (F) cytoskeletal reorganization, (G) reepithelialization, (H) unidentified role in wound healing, and (I) cholesterol biosynthesis. The numbers in (C) and (G) refer to genes whose products serve as signals to neutrophils (C1), monocytes and macrophages (C2), T lymphocytes (C3), B lymphocytes (C4), and melanocytes (G1).

from G$_0$ to a proliferating state. However, one of the defining characteristics of genome-scale expression profiling experiments is that the examination of so many diverse genes opens a window on all the processes that actually occur and not merely the single process one intended to observe. Serum, the soluble fraction of clotted blood, is normally encountered by cells in vivo in the context of a wound. Indeed, the expression program that we observed in response to serum suggests that fibroblasts are programmed to interpret the abrupt exposure to serum not as a general mitogenic stimulus but as a specific physiological signal, signifying a wound. The proliferative response that we originally intended to study appeared to be part of a larger physiological response of fibroblasts to a wound. Other features of the transcriptional response to serum suggest that the fibroblast is an active participant in a conversation among the diverse cells that work together in wound repair, interpreting, amplifying, modifying, and broadcasting signals controlling inflammation, angiogenesis, and epithelial regrowth during the response to an injury.

We recognize that these in vitro results almost certainly represent a distorted and incomplete rendering of the normal physiological response of a fibroblast to a wound. Moreover, only the responses elicited directly by exposure of fibroblasts to serum were examined. The subsequent signals from other cellular participants in the normal wound-healing process would certainly provoke further evolution of the transcriptional program in fibroblasts at the site of a wound, which this experiment cannot reveal. Nevertheless, we believe that the picture that emerged strongly suggests a much larger and richer role for the fibroblast in the orchestration of this important physiological process than had previously been suspected.

References and Notes

1. J. A. Winkles, *Prog. Nucleic Acid Res. Mol. Biol.* **58**, 41 (1998).
2. A normal human diploid fibroblast cell line derived from foreskin (ATCC CRL 2091) in passage 8 was used in these experiments. The protocol followed for growth arrest and stimulation was essentially that of (16) and (17). Cells were grown to about 60% confluence in 15-cm petri dishes in Dulbecco's minimum essential medium containing glucose (1 g/liter), the antibiotics penicillin and streptomycin, and 10% (by vol) FBS (Hyclone) that had been previously heat inactivated at 56°C for 30 min. The cells were then washed three times with the same medium lacking FBS, and low-serum medium (0.1% FBS) was added to the plates. After a 48-hour incubation, the medium was replaced with fresh medium containing 10% FBS. mRNA was isolated from several plates of cells harvested before serum stimulation; this mRNA served as the serum-starved or time-zero reference sample. Cells were harvested from batches of plates at 11 subsequent intervals (15 min, 30 min, 1, 2, 4, 6, 8, 12, 16, 20, and 24 hours) after the addition of serum. mRNA was also isolated from exponentially growing fibroblasts (not subjected to serum starvation). mRNA was isolated with the FastTrack mRNA isolation kit (Invitrogen), which involves lysis of the cells on the plate. The growth medium was removed, and the cells were quickly washed with phosphate-buffered saline at room temperature. The lysis buffer was added to the plate, transferred to tubes, and frozen in liquid nitrogen. Subsequent steps were performed according to the kit manufacturer's protocols.
3. The National Center for Biotechnology Information maintains the UniGene database as a resource for partitioning human sequences contained in GenBank into clusters representing distinct transcripts or genes (18, 19). At the time this work began, this database contained about 40,000 such clusters. We selected a subset of 10,000 of these UniGene clusters for inclusion on gene expression microarrays. UniGene clusters were included only if they contained at least one clone from the I.M.A.G.E. human cDNA collection (20), so that a physical clone could easily be obtained (all I.M.A.G.E. clones are available commercially from a number of vendors). We attempted to include as complete as possible a set of the "named" human genes (about 4000) and genes that appeared to be closely related to named genes in other organisms (about an additional 2000). The remaining 4000 clones were chosen from among the "anonymous" UniGene clusters on the basis of inclusion on the human transcript map (www.ncbi.nlm.nih.gov/SCIENCE96/) and the lack of apparent homology to any other genes in the selected set. A physical clone representing each of the selected genes was obtained from Research Genetics. This "10K set" is included in a more recent "15K set" described at www.nhgri.nih.gov/DIR/LCG/15K/HTML/p15Ktop.html. Of these clones, 472 are absent from the current edition of UniGene and were presumed to be distinct genes. The remainders represent 8141 distinct clusters, or human genes, in UniGene. These clones, thus presumed to represent 8613 different genes, were used to print microarrays according to methods described previously (21, 22).
4. One microgram of mRNA was used for making fluorescently labeled cDNA probes for hybridizing to the microarrays, with the protocol described previously (23). mRNA from the large batch of serum-starved cells was used to make cDNA labeled with Cy3. The Cy3-labeled cDNA from this batch of serum-starved cells served as the common reference probe in all hybridizations. mRNA samples from cells harvested immediately before serum stimulation, at intervals after serum stimulation, and from exponentially growing cells were used to make cDNA labeled with Cy5. Ten micrograms of yeast tRNA, 10 μg of polydeoxyadenylic acid, and 20 μg of human CoT1 DNA (Gibco-BRL) were added to the mixture of labeled probes in a solution containing 3× standard saline citrate (SSC) and 0.3% SDS and allowed to prehybridize at room temperature for 30 min before the probe was added to the surface of the microarray. Hybridizations, washes, and fluorescent scans were performed as described previously (23, 24). All measurements, totaling more than 180,000 differential expression measurements, were stored in a computer database for analysis and interpretation.
5. The nominal identities of a number of cDNAs (currently about 3750) on the microarray were verified by sequencing. The clones that were sequenced included many of the genes whose expression changed substantially upon serum stimulation, as well as a large number of genes whose expression did not change substantially in the course of this experiment. About 85% of the clones on the current version of this microarray that were checked by resequencing were correctly identified. In all the figures, gene names or EST numbers are given only for those genes on the microarrays whose identities were reconfirmed by resequencing. In the cases where a human gene has more than one name in the literature, we have tried to use the name that is most evocative of its presumed role in this context. The remainder of the clones have been assigned a temporary identification number (format: SID######) and a putative identity pending sequence verification. The correct identities of these genes will be posted at our Web site (genome-www.stanford.edu/serum) as they are confirmed by resequencing.
6. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
7. Genes were selected for this analysis if either (i) their expression level deviated from that in quiescent fibroblasts by at least a factor of 2.20 in at least two of the samples from serum-stimulated cells or (ii) the standard deviation for the set of 13 values of log$_2$(expression ratio) measured for the gene in this time course exceeded 0.7. In addition, observations in which the pixel-by-pixel correlation coefficients for the Cy3 and Cy5 fluorescence signals measured in a given array element were less than 0.6 were excluded. This selection criteria yielded a computationally manageable number of genes while minimizing the number of genes that were included because of noise in the data.
8. A more complete analysis and interpretation of the results of this experiment, as well as a searchable database, can be found at genome-www.stanford.edu/serum
9. K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, K. Deetz, *PCR Methods. Appl.* **4**, 357 (1995).
10. The apparent dip in the profile of COX2 at the 2-hour time point in the microarray data appears to result from a localized area of low intensity on the corresponding array scan resulting in an underestimation of the expression ratio. The expression ratios measured for mast/stem cell growth factor receptor are somewhat lower in the microarray data. This discrepancy is probably a consequence of the conservative background subtraction method used for quantitating the signal intensities on the array scans (23). The sequences of the PCR primer pairs (5' to 3') that were used are as follows: COX2, CCGTGGCTCTCTT-GGCAG and CTAAGTTCTTTAGCACTCCTTGGCA; IL-8, CGATGCTGTGGAGCTGTATC and CCATGGTTTC-ACCAAAGATG; mast/stem cell factor receptor, ACA-GAAGCCCGTGCTAGACC and GAGGCTGCGAGGAG-GAAG; B4-2, AAACCCCCCTCAGGAAAGAG and CC-ATGAACAAGCTGGCCAT; and actin, AGTACTCCGTGT-GGATCGGC and GCTGATCCACATCTGCTGCA
11. V. R. Iyer et al., unpublished data. The gene expression data for the early time points in the presence of cycloheximide will be available at our Web site (genome-www.stanford.edu/serum)
12. T. Hunter, *Cell* **80**, 225 (1995).
13. J. Leppaluoto and H. Ruskoaho, *Ann. Med.* **24**, 153 (1992).
14. A. C. Chang et al., *Mol. Cell. Endocrinol.* **112**, 241 (1995).
15. K. L. Madsen et al., *Am. J. Physiol.* **274**, G96 (1998).
16. W. Krek and J. A. DeCaprio, *Methods Enzymol.* **254**, 114 (1995).
17. R. A. Tobey, J. G. Valdez, H. A. Crissman, *Exp. Cell Res.* **179**, 400 (1988).
18. M. S. Boguski and G. D. Schuler, *Nature Genet.* **10**, 369 (1995).
19. G. D. Schuler, *J. Mol. Med.* **75**, 694 (1997).
20. G. Lennon, C. Auffray, M. Polymeropoulos, M. B. Soares, *Genomics* **33**, 151 (1996).
21. I.M.A.G.E. clones were amplified by PCR in 96-well format with amino-linked primers at the 5' end. Purified PCR products were suspended at a concentration of ~0.5 mg/ml in 3× SSC, and ~5 ng of each product was arrayed onto coated glass by means of procedures similar to those described previously (22). A total of 9996 elements were arrayed onto an area of 1.8 cm by 1.8 cm with the elements spaced 175 μm apart. The microarrays were then postprocessed to fix the DNA to the glass surface before hybridization with a procedure similar to previously described methods (22).
22. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
23. J. L. DeRisi, V. R. Iyer, P. O. Brown, *ibid.* **278**, 680 (1997).
24. J. DeRisi et al., *Nature Genet.* **14**, 457 (1996).
25. We thank E. Chung for help with sequencing, A. Alizadeh for help with sequence verification, K. Ranade for advice on the TaqMan assay, and J. DeRisi and other members of the P.O.B. and D.B. labs for discussions. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450) and the National Cancer Institute (NIH CA 77097). V.R.I. was supported in part by an Institutional Training Grant in Genome Sciences (T32 HG00044) from the NHGRI. M.B.E. is an Alfred E. Sloan Foundation Postdoctoral Fellow in Computational Molecular Biology, and D.T.R is a Walter and Idun Berry Fellow. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

13 August 1998; accepted 13 November 1998

Systematic variation in gene expression patterns in human cancer cell lines

Douglas T. Ross[1], Uwe Scherf[5], Michael B. Eisen[2], Charles M. Perou[2], Christian Rees[2], Paul Spellman[2], Vishwanath Iyer[1], Stefanie S. Jeffrey[3], Matt Van de Rijn[4], Mark Waltham[5], Alexander Pergamenschikov[2], Jeffrey C.F. Lee[6], Deval Lashkari[7], Dari Shalon[6], Timothy G. Myers[8], John N. Weinstein[5], David Botstein[2] & Patrick O. Brown[1,9]

W used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 c ll lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression patt rns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of th xpression patterns in the tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours *in vivo*.

Intr duction

Cell lines derived from human tumours have been extensively used as experimental models of neoplastic disease. Although such cell lines differ from both normal and cancerous tissue, the inaccessibility of human tumours and normal tissue makes it likely that such cell lines will continue to be used as experimental models for the foreseeable future. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines (the NCI60) derived from tumours from a variety of tissues and organs[1-4]. The DTP has assessed many molecular features of the cells related to cancer and chemotherapeutic sensitivity, and has measured the sensitivities of these 60 cell lines to more than 70,000 different chemical compounds, including all common chemotherapeutics (http://dtp.nci.nih.gov). A previous analysis of these data revealed a connection between the pattern of activity of a drug and its method of action. In particular, there was a tendency for groups of drugs with similar patterns of activity to have related methods of action[3,5-7].

We used DNA microarrays to survey the variation in abundance of approximately 8,000 distinct human transcripts in these 60 cell lines. Because of the logical connection between the function of a gene and its pattern of expression, the correlation of gene expression patterns with the variation in the phenotype of the cell can begin the process by which the function of a gene can be inferred. Similarly, the patterns of expression of known genes can reveal novel phenotypic aspects of the cells and tissues studied[8-10]. Here we present an analysis of the observed patterns of gene expression and their relationship to phenotypic properties of the 60 cell lines. The accompanying report[11] explores the relationship between the gene expression patterns and the drug sensitivity profiles measured by the DTP. The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions *in vitro* and *in vivo*, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes[11-16]. The databases, plus tools for analysis and visualization of the data, are available (http://genome-www.stanford.edu/nci60 and http://discover.nci.nih.gov).

Results

We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides[17,18]. The cDNAs included approximately 8,000 different genes: approximately 3,700 represented previously characterized human proteins, an additional 1,900 had homologues in other organisms and the remaining 2,400 were identified only by ESTs. Due to ambiguity of the identity of the cDNA clones used in these studies, we estimated that approximately 80% of the genes in these experiments were correctly identified. The identities of approximately 3,000 cDNAs

Departments of [1]Biochemistry, [2]Genetics, [3]Surgery and [4]Pathology, Stanford University School of Medicine, Stanford, California, USA. [5]Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [6]Incyte Pharmaceuticals, Fremont, California, USA. [7]Genometrix Inc., The Woodlands, Texas, USA. [8]Information Technology Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA. [9]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu) or J.N.W. (e-mail: Weinstein@dtpax2.ncifcrf.gov).
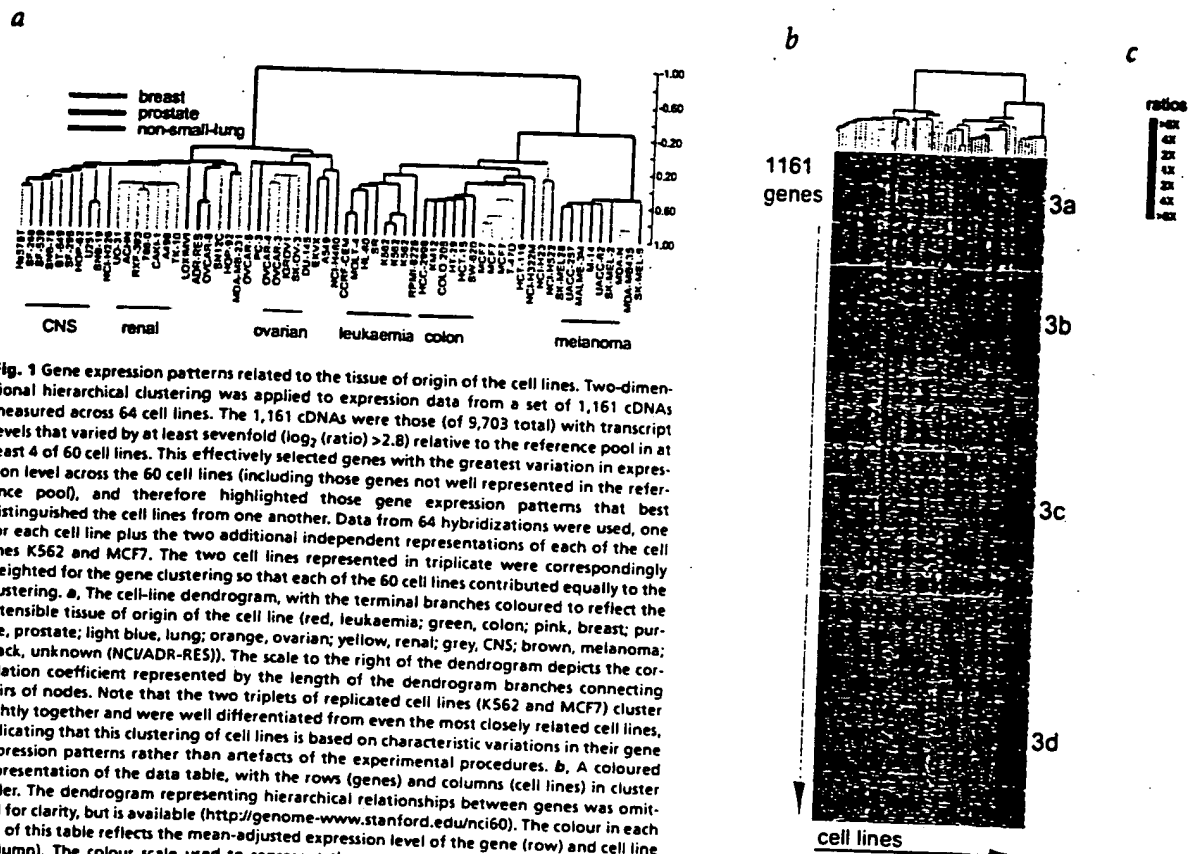
a

b

c

Fig. 1 Gene expression patterns related to the tissue of origin of the cell lines. Two-dimensional hierarchical clustering was applied to expression data from a set of 1,161 cDNAs measured across 64 cell lines. The 1,161 cDNAs were those (of 9,703 total) with transcript levels that varied by at least sevenfold ($\log_2$ (ratio) >2.8) relative to the reference pool in at least 4 of 60 cell lines. This effectively selected genes with the greatest variation in expression level across the 60 cell lines (including those genes not well represented in the reference pool), and therefore highlighted those gene expression patterns that best distinguished the cell lines from one another. Data from 64 hybridizations were used, one for each cell line plus the two additional independent representations of each of the cell lines K562 and MCF7. The two cell lines represented in triplicate were correspondingly weighted for the gene clustering so that each of the 60 cell lines contributed equally to the clustering. *a*, The cell-line dendrogram, with the terminal branches coloured to reflect the ostensible tissue of origin of the cell line (red, leukaemia; green, colon; pink, breast; purple, prostate; light blue, lung; orange, ovarian; yellow, renal; grey, CNS; brown, melanoma; black, unknown (NCI/ADR-RES)). The scale to the right of the dendrogram depicts the correlation coefficient represented by the length of the dendrogram branches connecting pairs of nodes. Note that the two triplets of replicated cell lines (K562 and MCF7) cluster tightly together and were well differentiated from even the most closely related cell lines, indicating that this clustering of cell lines is based on characteristic variations in their gene expression patterns rather than artefacts of the experimental procedures. *b*, A coloured representation of the data table, with the rows (genes) and columns (cell lines) in cluster order. The dendrogram representing hierarchical relationships between genes was omitted for clarity, but is available (http://genome-www.stanford.edu/nci60). The colour in each cell of this table reflects the mean-adjusted expression level of the gene (row) and cell line (column). The colour scale used to represent the expression ratios is shown. The labels '3a–3d' in (*b*) refer to the clusters of genes shown in detail in Fig. 3.

from these experiments have been sequence-verified, including all of those referred to here by name.

Each hybridization compared Cy5-labelled cDNA reverse transcribed from mRNA isolated from one of the cell lines with Cy3-labelled cDNA reverse transcribed from a reference mRNA sample. This reference sample, used in all hybridizations, was prepared by combining an equal mixture of mRNA from 12 of the cell lines (chosen to maximize diversity in gene expression as determined primarily from two-dimensional gel studies[2]). By comparing cDNA from each cell line with a common reference, variation in gene expression across the 60 cell lines could be inferred from the observed variation in the normalized Cy5/Cy3 ratios across the hybridizations.

To assess the contribution of artefactual sources of variation in the experimentally measured expression patterns, K562 and MCF7 cell lines were each grown in three independent cultures, and the entire process was carried out independently on mRNA extracted from each culture. The variance in the triplicate fluorescence ratio measurements approached a minimum when the fluorescence signal was greater than approximately 0.4% of the measurable total signal dynamic range above background in either channel of the hybridization. We selected the subset of spots for which significant signal was present in both the numerator and denominator of the ratios by this criterion to identify the best-measured spots. The pair-wise correlation coefficients for the triplicates of the set of genes that passed this quality control level (6,992 spots included for the MCF7 samples and 6,161 spots for K562) ranged from 0.83 to 0.92 (for graphs and details, see http://genome-www.stanford.edu/nci60).

To make the orderly features in the data more apparent, we used a hierarchical clustering algorithm[19,20] and a pseudo-colour visu-

alization matrix[3,21]. The object of the clustering was to group cell lines with similar repertoires of expressed genes and to group genes whose expression level varied among the 60 cell lines in a similar manner. Clustering was performed twice using different subsets of genes to assess the robustness of the analysis. In one case (Fig. 1), we concentrated on those genes that showed the most variation in expression among the 60 cell lines (1,167 total). A second analysis (Fig. 2) included all spots that were thought to be well measured in the reference set (6,831 spots).

Gene expression patterns related to the histologic origins of the cell lines

The most notable property of the clustered data was that cell lines with common presumptive tissues of origin grouped together (Figs 1*a* and 2). Cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Many of these coherent cell line clusters were distinguished by the specific expression of characteristic groups of genes (Fig. 3*a–d*). For example, a cluster of approximately 90 genes was highly expressed in the melanoma-derived lines (Fig. 3*c*). This set was enriched for genes with known roles in melanocyte biology, including tyrosinase and dopachrome tautomerase (TYR and DCT; two subunits of an enzyme complex involved in melanin synthesis[22]), MART1 (MLANA; which is being investigated as a target for immunotherapy of melanoma[23]) and S100-β (S100B; which has been used as an antigenic marker in the diagnosis of

Fig. 2 Gene expression patterns related to other cell-line phenotypes. **a,** We applied two-dimensional hierarchical clustering to expression data from a set of 6,831 cDNAs measured across the 64 cell lines. The 6,831 cDNAs were those with a minimum fluorescence signal intensity of approximately 0.4% of the dynamic range above background in the reference channel in each of the six hybridizations used to establish reproducibility. This effectively selected those spots that provided the most reliable ratio measurements and therefore identified a subset of genes useful for exploring patterns comprised of those whose variation in expression across the 60 cell lines was of moderate magnitude. **b,** Cluster-ordered data table. **c,** Doubling time of cell lines. Cell lines are given in cluster order. Values are plotted relative to the mean. Doubling times greater than the mean are shown in green, those with doubling time less than the mean are shown in red. **d,** Three related gene clusters that were enriched for genes whose expression level variation was correlated with cell line proliferation rate. Each of the three gene clusters (clustered solely on the basis of their expression patterns) showed enrichment for sets of genes involved in distinct functional categories (for example, ribosomal genes versus genes involved in pre-RNA splicing). **e,** Gene cluster in which all characterized and sequence-verified cDNAs encode genes known to be regulated by interferons. **f,** Gene cluster enriched for genes that have been implicated in drug metabolism (indicated by asterisks). A further property of the gene clustering evident here and in Fig. 2 is the strong tendency for redundant representations of the same gene to cluster immediately adjacent to one another, even within larger groups of genes with very similar expression patterns. In addition to illustrating the reproducibility and consistency of the measurements, and providing independent confirmation of many of our measurements, this property also demonstrates that these, and probably all, genes have nearly unique patterns of variation across the 60 cell lines. If this were not the case, and multiple genes had identical patterns of variation, we would not expect to be able to distinguish, by clustering on the basis of expression variation, duplicate copies of individual genes from the other genes with identical expression patterns.

d proliferation cluster

e interferon cluster

f drug metabolism cluster

cell lines

melanoma). LOXIMVI, the seventh line designated as melanoma in the NCI60, did not show this characteristic pattern. Although isolated from a patient with melanoma, LOXIMVI has previously been n ted t lack melanin and other markers useful for identification of melanoma cells[1].

Paradoxically, two related cell lines (MDA-MB435 and MDA-N), which were derived from a single patient with breast cancer and have been conventionally regarded as breast cancer cell lines, shared expression of the genes associated with melanoma. MDA-MB435 was isolated from a pleural effusion in a patient with metastatic ductal adenocarcinoma of the breast[24,25]. It remains possible that the origin of the cell line was a breast cancer, and that its gene expression pattern is related to the neuroendocrine features of some breast cancers[26]. But our results suggest that this cell line may have originated from a melanoma, raising the possibility that the patient had a co-existing occult melanoma.

The higher-level organization of the cell-line tree—in which groups span cell lines from different tissue types—also reflected shared biological properties of the tissues from which the cell lines were derived. The carcinoma-derived cell lines were divided into major branches that separated those that expressed genes characteristic of epithelial cells from those that expressed genes more typical of stromal cells. A cluster of genes is shown (Fig. 3b) that is most strongly expressed in cell lines derived from colon carcinomas, six of seven ovarian-derived cell lines and the two breast cancer lines positive for the oestrogen receptor. The named genes in this cluster have been implicated in several aspects of epithelial cell biology[27]. The cluster was enriched for genes whose products are known to localize to the basolateral membrane of epithelial cells, including those encoding components of adherens complexes (for example, desmoplakin (DSP), periplakin (PPL) and plakoglobin (JUP)), an epithelial-expressed cell-cell adhesion molecule (M4S1) and a sodium/hydrogen ion exchanger[28–31] (SLC9A1). It also contained genes that encode putative transcriptional regulators of epithelial morphogenesis, a human homologue of a *Drosophila melanogaster* epithelial-expressed tumour suppressor (LLGL1) and a homeobox gene thought to control calcium-mediated adherence in epithelial cells[32,33] (MSX2).

In contrast, a separate, major branch of the cell-line dendrogram (Fig. 1a) included all glioblastoma-derived cell lines, all renal-cell-carcinoma–derived cell lines and the remaining carcinoma-derived lines. The characteristic set of genes expressed in this cluster included many whose products are involved in stromal cell functions (Fig. 3d). Indeed, the two cell lines originally described as 'sarcoma-like' in appearance (Hs578T, breast carcinosarcoma, and SF539, gliosarcoma) expressed most of these genes[34,35]. Although no single gene was uniformly characteristic of this cluster, each cell line showed a distinctive pattern of expression of genes encoding proteins with roles in synthesis or modification of the extracellular matrix (for example, caldesmon (CALD1), cathepsins, thrombospondin (THBS), lysyl oxidase (LOX) and collagen subtypes). Although the ovarian and most non-small-cell-lung–derived carcinomas expressed genes characteristic of both epithelial cells and stromal cells, they probably clustered with the CNS and renal cell carcinomas in this analysis because genes characteristically expressed in stromal cells were more abundantly represented in this gene set.

Physiological variation reflect d in g ne expressi n patterns

A cluster diagram of 6,831 genes (Fig. 2) is useful for exploring clusters of genes whose variation in mRNA levels was not obviously attributable to cell or tissue type. We identified some gene clusters that were enriched for genes involved in specific cellular

processes; the variation in their expressi n levels may reflect corresponding differences in activity of these processes in the cell lines. For example, a cluster f 1,159 genes (Fig. 2a) included many whose products are necessary f r progression through the cell cycle (such as CCNA1, MCM106 and MAD2L1), RNA processing and translation machinery (such as RNA helicases, hnRNPs and translation elongation factors) and traditional pathologic markers used to identify proliferating cells (MKI67). Within this large cluster were smaller clusters enriched for genes with more specialized roles. One cluster was highly enriched for numerous ribosomal genes, whereas another was more enriched for genes encoding RNA-splicing factors. The variation in expression of these ribosomal genes was significantly correlated with variation in the cell doubling time (correlation coefficient of 0.54), supporting the notion that the genes in this cluster were regulated in relation to cell proliferation rate or growth rate in these cell lines.
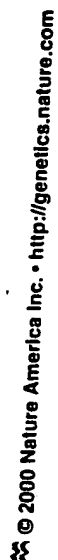
In a smaller gene cluster (Fig. 2d), all of the named genes were previously known to be regulated by interferons[13,36]. Additional groups of interferon-regulated genes showed distinct patterns of expression (data not shown), suggesting that the NCI60 cell lines exhibited variation in activity of interferon-response pathways, which was reflected in gene expression patterns[36].

Another cluster (Fig. 2e) contained several genes encoding proteins with possible interrelated roles in drug metabolism, including glutamate-cysteine ligase (GLCLC, the enzyme responsible for the rate limiting step of glutathione synthesis), thioredoxin (TXN) and thioredoxin reductase (TXNRD1; enzymes involved in regulating redox state in cells), and MRP1 (a drug transporter known to efficiently transport glutathione-conjugated compounds[37]). The elevated expression of this set of genes in a subset of these cell lines may reflect selection for resistance to chemotherapeutics.

Cell lines facilitate interpretation of gene expression patterns in complex clinical samples

Like many other types of cancer, tumours of the breast typically have a complex histological organization, with connective tissue and leukocytic infiltrates interwoven with tumour cells. To explore the possibility that variation in gene expression in the tumour cell lines might provide a framework for interpreting the expression patterns in tumour specimens, we compared RNA isolated from two breast cancer biopsy samples, a sample of normal breast tissue and the NCI60 cell lines derived from breast cancers (excluding MDA-MB-435 and MDA-N) and leukaemias (Fig. 4). This clustering highlighted features of the gene expression pattern shared between the cancer specimens and individual cell lines derived from breast cancers and leukaemias.

The genes encoding keratin 8 (KRT8) and keratin 19 (KRT19), as well as most of the other 'epithelial' genes defined in the complete NCI60 cell line cluster, were expressed in both of the biopsy samples and the two breast-derived cell lines, MCF-7 and T47D, expressing the oestrogen receptor, suggesting that these transcripts originated in tumour cells with features similar to those of luminal epithelial cells (Fig. 5a). Expression of a set of genes characteristic of stromal cells, including collagen genes (COL3A1, COL5A1 and COL6A1) and smooth muscle cell markers (TAGLN), was a feature shared by the tumour sample and the stromal-like cell lines Hs578T and BT549 (Fig. 5b). This feature of the expression pattern seen in the tumour samples is likely to be due to the stromal component of the tumour. The tumours also shared expression of a set of genes (Fig. 5c) with the multiple myeloma cell line (RPMI-8226), notably including immunoglobulin genes, consistent with the presence of B cells in the tumour (this was confirmed by staining with anti-

230

a

leukaemia cluster (6 ESTs)

b

epithelial cluster (15 ESTs)

c

melanoma cluster (16 ESTs)

d

mesenchymal cluster (67 ESTs)

Fig. 3 Gene clusters related to tissue characteristics in the cell lines. Enlargements of the regions of the cluster diagram in Fig. 1 showing gene clusters enriched for genes expressed in cell lines of ostensibly similar origins. *a*, Cluster of genes highly expressed in the leukaemia-derived cell lines. Two sub-clusters distinguish genes that were expressed in most leukaemia-derived lines from those expressed exclusively in the eryroblastoid line, K562 (note that the triplicate hybridizations cluster together). *b*, Cluster of genes highly expressed in all colon (7/7) cell lines and all breast-derived cell lines positive for the oestrogen receptor (2/2). This set of genes was also moderately expressed in most ovarian lines (5/6) and some non-small-cell-lung (4/6) lines, but was expressed at a lower level in all renal-cancer–derived lines. *c*, Cluster of genes highly expressed in most melanoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB435 and MDA-N). *d*, Cluster of genes highly expressed in all glioblastoma (6/6) lines and most lines derived from renal-cell carcinoma (7/8), and more moderately expressed in a subset of carcinoma-derived lines. In all panels, names are shown only for all known genes whose identities were independently re-verified by sequencing. The number of sequence-validated ESTs within the cluster is indicated below the cluster in parentheses. The position of gene names in the adjacent list only approximates their position in the cluster diagram as indicated by the lines connecting the colour chart with the gene list. Complete cluster images with all gene names and accession numbers are available (http://genome-www.stanford.edu/nci60).
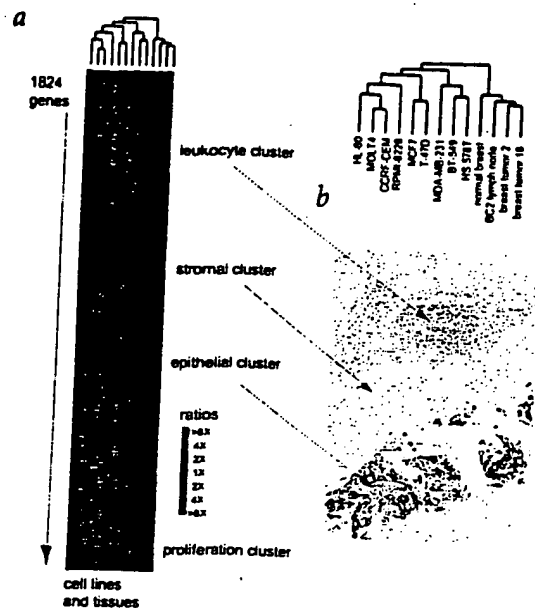
Fig. 4 Comparison of the gene expression patterns in clinical breast cancer specimens and cultured breast cancer and leukaemia cell lines. *a*, Two-dimensional hierarchical clustering applied to gene expression data for two breast cancer specimens, a lymph node metastasis from one patient, normal breast and the NCI60 breast and leukaemia-derived cell lines. The gene expression data from tissue specimens was clustered along with expression data from a subset of the NCI60 cell lines to explore whether features of expression patterns observed in specific lines could be identified in the tissue samples. Labels indicate gene clusters (shown in detail in Fig. 5) that may be related to specific cellular components of the tumour specimens. *b*, Breast cancer specimen 16 stained with anti-keratin antibodies, showing the complex mix of cell types characteristically found in breast tumours. The arrows highlight the different cellular components of this tissue specimen that were distinguished by the gene expression cluster analysis (Fig. 5).

---

immunoglobulin antibodies; data not shown). Therefore, distinct sets of genes with co-varying expression among the samples (Fig. 4, arrow) appear to represent distinct cell types that can be distinguished in breast cancer tissue. A fourth cluster of genes, more highly expressed in all of the cell lines than in any of the clinical specimens, was enriched for genes present in the 'proliferation' cluster described above (Fig. 5*d*). The variation in expression of these genes likely paralleled the difference in proliferation rate between the rapidly cycling cultured cell lines and the much more slowly dividing cells in tissues.

Discussion

Newly available genomics tools allowed us to explore variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumour tissues. We used a simple cluster analysis to identify the prominent features in the gene expression patterns that appeared to reflect 'molecular signatures' of the tissue from which the cells originated. The histological characteristics of the cell lines that dominated the clustering were pervasive enough that similar relationships were revealed when alternative subsets of genes were selected for analysis. Additional features of the expression pattern may be related to variation in physiological attributes such as proliferation rate and activity of interferon-response pathways.

The properties of the tumour-derived cell lines in this study have presumably all been shaped by selection for resistance to host defences and chemotherapeutics and for rapid proliferation in the tissue culture environment of synthetic growth media, fetal bovine serum and a polystyrene substratum. But the primary identifiable factor accounting for variation in gene expression patterns among these 60 cell lines was the identity of the tissue from which each cell line was ostensibly derived. For most of the cell lines we examined, neither physiological nor experimental adaptation for growth in culture was sufficient to overwrite the gene expression programs established during differentiation *in vivo*. Nevertheless, the prominence of mesenchymal features in the cell lines isolated from glioblastomas and carcinomas may reflect a selection for the relative ease of establishment of cell lines expressing stromal characteristics, perhaps combined with physiological adaptation to tissue culture conditions[38–40].

Biological themes linking genes with related expression patterns may be inferred in many cases from the shared attributes of known genes within the clusters. Uncharacterized cDNAs are likely to encode proteins that have roles similar to those of the known gene products with which they appear to be co-regulated. Still, for several clusters of genes, we were unable to discern a common theme linking the identified members of the cluster. Further exploration of their variation in expression under more diverse conditions and more comprehensive investigation of the physiology of the NCI60 cells may provide insight[10]. The relationship of the gene expression patterns to the drug sensitivity patterns measured by the DTP is an example of linking variation in gene expression with more subtle and diverse phenotypic variation[11].

The patterns of gene expression measured in the NCI60 cell lines provide a framework that helps to distinguish the cells that express specific sets of genes in the histologically complex breast cancer specimens[41]. Although it is now feasible to analyse gene expression in micro-dissected tumour specimens[42,43], this observation suggests that it will be possible to explore and interpret some of the biology of clinical tumour samples by sampling them intact. As is useful in conventional morphological pathology, one might be able to observe interactions between a tumour and its microenvironment in this way. These relationships will be clarified by suitable analysis of gene expression patterns from intact as well as dissected tumours[12,14,15,41].

Methods

(100 µl) was purified by gel exclusion, concentrated and resuspended in 3×SSC (10 µl). The PCR products were then printed on treated glass microscope slides using a robot with four printing tips. Detailed protocols for assembling and operating a microarray printer, and printing and experimental application of DNA microarrays are available (http://cmgm. stanford.edu/pbrown).

Preparation of mRNA and reference pool. Cell lines were grown from NCI DTP frozen stocks in RPMI-1640 supplemented with phenol red, glutamine (2 mM) and 5% fetal calf serum. To minimize the contribution of variations in culture conditions or cell density to differential gene expression, we grew each cell line to 80% confluence and isolated mRNA 24 h after transfer to fresh medium. The time between removal from the incubator and lysis of the cells in RNA stabilization buffer was minimized (<1 min). Cells were lysed in buffer containing guanidium isothiocyanate and total RNA was purified with the RNeasy purification kit (Qiagen). We purified mRNA as needed

using a poly(A) purification kit (Oligotex, Qiagen) according to the manufacturer's instructions. Denaturing agarose gel electrophoresis assessed the integrity and relative contamination of mRNA with ribosomal RNA.

The breast tumours were surgically excised from patients and rapidly transported to the pathology laboratory, where samples for microarray analysis were quickly frozen in liquid nitrogen and stored at –80 °C until use. A frozen tumour specimen was removed from the freezer, cut into small pieces (~50–100 mg each), immediately placed into 10–12 ml of Trizol reagent (Gibco-BRL) and homogenized using a PowerGen 125 Tissue Homogenizer (Fisher Scientific), starting at 5,000 r.p.m. and gradually increasing to ~20,000 r.p.m. over a period of 30–60 s. We processed the Trizol/tumour homogenate as described in the Trizol protocol, including an initial step to remove fat. Once total RNA was obtained, we isolated mRNA with a FastTrack 2.0 kit (Invitrogen) using the manufacturer's protocol for isolating mRNA starting from total RNA. The normal breast samples were obtained from Clontech.

Fig. 5 Histologic features of breast cancer biopsies can be recognized and parsed based on gene expression patterns. Enlargements of the regions of the cluster diagram in Fig. 4 showing gene clusters enriched for genes expressed in different cell types in the breast cancer specimens, as distinguished by clustering with the cultured cell lines. *a*, A cluster including many genes characteristic of epithelial cells expressed in cell lines (T47D and MCF7) derived from breast cancer positive for the oestrogen receptor and tumours. *b*, Genes expressed in cell lines derived from breast cancer with stromal cell characteristics (Hs578T and BT549) and tumour specimens. Expression of these genes in the tumour samples may reflect the presence of myofibroblasts in the cancer specimen stroma. *c*, Genes expressed in leukocyte-derived cell lines, showing common leukocyte, and separate 'myeloid' and 'B-cell', gene clusters. *d*, Genes that were relatively highly expressed in all cell lines compared with the tumour specimens and normal breast. The higher expression of this set of genes involved in cell cycle transit in the cell lines is likely to reflect the higher proliferative rate of cells cultured in the presence of serum compared with the average proliferation rate of cells in the biopsied tissue.

We combined mRNA from the following cells in equal quantities to make the reference pool: HL-60 (acute myeloid leukaemia) and K562 (chronic myeloid leukaemia); NCI-H226 (non-small-cell-lung); COLO 205 (colon); SNB-19 (central nervous system); LOX-IMVI (melanoma); OVCAR-3 and OVCAR-4 (ovarian); CAKI-1 (renal); PC-3 (prostate); and MCF7 and Hs578T (breast). The criterion for selection of the cell lines in the reference are described in detail in the accompanying manuscript[12].

Doubling-time calculations. We calculated doubling times based on routine NCI60 cell line compound screening data; and they reflect the doubling times for cells inoculated into 96-well plates at the screening inoculation densities and grown in RPMI 1640 medium supplemented with 5% fetal bovine serum for 48 h. We measured cell populations using sulforhodamine B optical density measurement assay. The doubling time constant k was calculated using the equation: $N/No = e^{kt}$, where No is optical density for control (untreated) cells at time zero, N is optical density for control cells after 48-h incubation, and t is 48 h. The same equation was then used with the derived k to calculate the doubling time t by setting $N/No = 2$. For a given cell line, we obtained No and N values by averaging optical densities (N>6,000) obtained for each cell line for a year's screening. Data and experimental details are available (http://dtp.nci.nih.gov).

Preparation and hybridization of fluorescent labelled cDNA. For each comparative array hybridization, labelled cDNA was synthesized by reverse transcription from test cell mRNA in the presence of Cy5-dUTP, and from the reference mRNA with Cy3-dUTP, using the Superscript II reverse-transcription kit (Gibco-BRL). For each reverse transcription reaction, mRNA (2 µg) was mixed with an anchored oligo-dT (d-20T-d(AGC)) primer (4 µg) in a total volume of 15 µl, heated to 70 °C for 10 min and cooled on ice. To this sample, we added an unlabelled nucleotide pool (0.6 µl; 25 mM each dATP, dCTP, dGTP, and 15 mM dTTP), either Cy3 or Cy5 conjugated dUTP (3 µl; 1 mM; Amersham), 5×first-strand buffer (6 µl; 250 mM Tris-HCL, pH 8.3, 375 mM KCl, 15 mM MgCl₂), 0.1 M DTT (3 µl) and 2 µl of Superscript II reverse transcriptase (200 µl/µl). After a 2-h incubation at 42 °C, the RNA was degraded by adding 1 N NaOH (1.5 µl) and incubating at 70 °C for 10 min. The mixture was neutralized by adding of 1 N HCL (1.5 µl), and the volume brought to 500 µl with TE (10 mM Tris, 1 mM EDTA). We added Cot1 human DNA (20 µg; Gibco-BRL), and purified the probe by centrifugation in a Centricon-30 micro-concentrator (Amicon). The two separate probes were combined, brought to a volume of 500 µl, and concentrated again to a volume of less than 7 µl. We added 10 µg/µl poly(A) RNA (1 µl; Sigma) and tRNA (10 µg/µl; Gibco-BRL) were added, and adjusted the volume to 9.5 µl with distilled water. For final probe preparation, 20×SSC (2.1 µl; 1.5 M NaCl, 150 mM NaCitrate, pH 8.0) and 10% SDS (0.35 µl) were added to a total final volume of 12 µl. The probes were denatured by heating for 2 min at 100 °C, incubated at 37 °C for 20–30 min, and placed on the array under a 22 mm×22 mm glass coverslip. We incubated slides overnight at 65 °C for 14–18 h in a custom slide chamber with humidity maintained by a small reservoir of 3×SSC. Arrays were washed by submersion and agitation for 2–5 min in 2×SSC with 0.1% SDS, followed by 1×SSC and then 0.1×SSC. The arrays were "spun dry" by centrifugation for 2 min in a slide-rack in a Beckman GS-6 tabletop centrifuge in Microplus carriers at 650 r.p.m. for 2 min.

Array quantitation and data processing. Following hybridization, arrays were scanned using a laser-scanning microscope (ref. 17; http://cmgm.stanford.edu/pbrown). Separate images were acquired for Cy3 and Cy5. We carried out data reduction with the program ScanAlyze (M.B.E., available

at http://rana.stanford.edu/software). Each spot was defined by manual positioning of a grid of circles over the array image. For each fluorescent image, the average pixel intensity within each circle was determined, and a local background was computed for each spot equal to the median pixel intensity in a square of 40 pixels in width and height centred on the spot centre, excluding all pixels within any defined spots. Net signal was determined by subtraction of this local background from the average intensity for each spot. Spots deemed unsuitable for accurate quantitation because of array artefacts were manually flagged and excluded from further analysis. Data files generated by ScanAlyze were entered into a custom database that maintains web-accessible files. Signal intensities between the two fluorescent images were normalized by applying a uniform scale factor to all intensities measured for the Cy5 channel. The normalization factor chosen so that the mean log(Cy3/Cy5) for a subset of spots that achieved a minimum quality parameter (approximately 6,000 spots) was 0. This effectively defined the signal-intensity-weighted 'average' spot on each array to have a Cy3/Cy5 ratio of 1.0.

Cluster analysis. We extracted tables (rows of genes, columns of individual microarray hybridizations) of normalized fluorescence ratios from the database. Various selection criteria, discussed in relation to each data set, were applied to select subsets of genes from the 9,703 cDNA elements on the arrays. Before clustering and display, the logarithm of the measured fluorescence ratios for each gene were centred by subtracting the arithmetic mean of all ratios measured for that gene. The centring makes all subsequent analyses independent of the amount of each gene's mRNA in the reference pool.

We applied a hierarchical clustering algorithm separately to the cell lines and genes using the Pearson correlation coefficient as the measure of similarity and average linkage clustering[3,19–21]. The results of this process are two dendrograms (trees), one for the cell lines and one for the genes, in which very similar elements are connected by short branches, and longer branches join elements with diminishing degrees of similarity. For visual display the rows and columns in the initial data table were reordered to conform to the structures of the dendrograms obtained from the cluster analysis. Each cell in the cluster-ordered data table was replaced by a graded colour (pure red through black to pure green), representing the mean-adjusted ratio value in the cell. Gene labels in cluster diagrams are displayed here only for genes that were represented in the microarray by sequence-verified cDNAs. A complete software implementation of this process is available (http://rana.stanford.edu/software), as well as all clustering results (http://genome-www.stanford.edu/nci60).

article

1. Stinson, S.F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* **12**, 1035–1053 (1992).
2. Myers, T.G. *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* **18**, 647–653 (1997).
3. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* **275**, 343–349 (1997).
4. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* **12**, 533–541 (1997).
5. Paull, K.D. *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl Cancer Inst.* **81**, 1088–1092 (1989).
6. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* **258**, 447–451 (1992).
7. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* **86**, 1853–1859 (1994).
8. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680–686 (1997).
9. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* **283**, 83–87 (1999).
10. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* **21** (suppl.), 33–37 (1999).
11. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* **24**, 236–244 (2000).
12. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
13. Der, S.D., Zhou, A., Williams, B.R. & Silverman, R.H. Identification of genes differentially regulated by interferon-α, -β or -γ or using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **95**, 15623–15628 (1998).
14. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* **96**, 6745–6750 (1999).
15. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* **229**, 101–108 (1999).
16. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).
17. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* **6**, 639–645 (1996).
18. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* **303**, 179–205 (1999).
19. Sokal, R.R. & Sneath, P.H.A. *Principles of Numerical Taxonomy* (W.H. Freeman, San Francisco, 1963).
20. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).
22. del Marmol, V. & Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* **381**, 165–168 (1996).
23. Kawakami, Y. *et al.* The use of melanosomal proteins in the immunotherapy of melanoma. *J. Immunother.* **21**, 237–246 (1998).
24. Cailleau, R., Olive, M. & Cruciger, Q.V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* **14**, 911–915 (1978).
25. Brinkley, B.R. *et al.* Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. *Cancer Res.* **40**, 3118–3129 (1980).
26. Nesland, J.M., Holm, R., Johannessen, J.V. & Gould, V.E. Neuroendocrine differentiation in breast lesions. *Pathol. Res. Pract.* **183**, 214–221 (1988).
27. Davies, J.A. & Garrod, D.R. Molecular aspects of the epithelial phenotype. *Bioessays* **19**, 699–704 (1997).
28. Garrod, D., Chidgey, M. & North, A. Desmosomes: differentiation, development, dynamics and disease. *Curr. Opin. Cell Biol.* **8**, 670–678 (1996).
29. Cowin, P. & Burke, B. Cytoskeleton-membrane interactions. *Curr. Opin. Cell Biol.* **8**, 56–65 (1996); erratum: **8**, 244 (1996).
30. Litvinov, S.V. *et al.* Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *J. Cell Biol.* **139**, 1337–1348 (1997).
31. Helmle-Kolb, C. *et al.* Na/H exchange activities in NHE1-transfected OK-cells: cell polarity and regulation. *Pflugers Arch.* **425**, 34–40 (1993); erratum: **427**, 387 (1994).
32. Manfruelli, P., Arquier, N., Hanratty, W.P. & Semeriva, M. The tumor suppressor gene, lethal(2)giant larvae (1(2)g1), is required for cell shape change of epithelial cells during Drosophila development. *Development* **122**, 2283–2294 (1996).
33. Lincecum, J.M., Fannon, A., Song, K., Wang, Y. & Sassoon, D.A. Msh homeobox genes regulate cadherin-mediated cell adhesion and cell-cell sorting. *J. Cell Biochem.* **70**, 22–28 (1998).
34. Hackett, A.J. *et al.* Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines. *J. Natl Cancer Inst.* **58**, 1795–1806 (1977).
35. Rutka, J.T. *et al.* Establishment and characterization of a cell line from a human gliosarcoma. *Cancer Res.* **46**, 5893–5902 (1986).
36. Nguyen, H., Hiscott, J. & Pitha, P.M. The growing family of interferon regulatory factors. *Cytokine Growth Factor Rev.* **8**, 293–312 (1997).
37. Moscow, J.A., Schneider, E., Ivy, S.P. & Cowan, K.H. Multidrug resistance. *Cancer Chemother. Biol. Response Modif.* **17**, 139–177 (1997).
38. Smith, H.S. & Hackett, A.J. The use of cultured human mammary epithelial cells in defining malignant progression. *Ann. N Y Acad. Sci.* **464**, 288–300 (1986).
39. Rutka, J.T. *et al.* Establishment and characterization of five cell lines derived from human malignant gliomas. *Acta Neuropathol.* **75**, 92–103 (1987).
40. Ronnov-Jessen, L., Petersen, O.W. & Bissell, M.J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* **76**, 69–125 (1996).
41. Perou, C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* **96**, 9212–9217 (1999).
42. Bonner, R.F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* **278**, 1481–1483 (1997).
43. Sgroi, D.C. *et al.* In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* **59**, 5656–5661 (1999).

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68	A1	(11) International Publication Number: **WO 95/21944**
		(43) International Publication Date: 17 August 1995 (17.08.95)

(21) International Application Number: PCT/US95/01863

(22) International Filing Date: 14 February 1995 (14.02.95)

(30) Priority Data:
08/195,485   14 February 1994 (14.02.94)   US

(60) Parent Application or Grant
(63) Related by Continuation
US   08/195,485 (CIP)
Filed on   14 February 1994 (14.02.94)

(71) Applicant *(for all designated States except US)*: SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ROSENBERG, Martin [US/US]; 241 Mingo Road, Royersford, PA 19468 (US). DEBOUCK, Christine [BE/US]; 667 Pugh Road, Wayne, PA 19087 (US). BERGSMA, Derk [US/US]; 271 Irish Road, Berwyn, PA 19312 (US).

(74) Agents: JERVIS, Herbert, H. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: DIFFERENTIALLY EXPRESSED GENES IN HEALTHY AND DISEASED SUBJECTS

(57) Abstract

The present invention involves methods and compositions for identifying genes which are differentially expressed in a normal healthy animal and an animal having a selected disease or infection, and methods for diagnosing diseases or infections characterized by the presence of those genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST isolated from an identified DNA library prepared from tissue or cell samples of a healthy animal, an animal with a selected disease or infection, and any combination thereof. Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of disease based on differential expression of genes of unknown function, and enable the identification of those genes and the proteins encoded thereby.

differentially expressed genes in healthy and diseased subjects

5   This application is a continuation-in-part application of U.S. Serial No.
08/195,485 filed February 14, 1994, the contents of which are incorporated herein by
reference.

Field of the Invention

10   The present invention relates to the use of immobilized
oligonucleotide/polynucleotide or polynucleotide sequences for the identification,
sequencing and characterization of genes which are implicated in disease, infection,
or development and the use of such identified genes and the proteins encoded thereby
in diagnosis, prognosis, therapy and drug discovery.

15

Background of the Invention

           Identification, sequencing and characterization of genes, especially
human genes, is a major goal of modern scientific research. By identifying genes,
determining their sequences and characterizing their biological function, it is possible
20   to employ recobinant DNA technology to produce large quantities of valuable "gene
products", e.g., proteins and peptides. Additionally, knowledge of gene sequences
can provide a key to diagnosis, prognosis and treatment of a variety of disease states
in plants and animals which are characterized by inappropriate expression and/or
repression of selected gene(s) or by the influence of external factors, e.g., carcinogens
25   or teratogens, on gene function. The term disease-associated genes(s) is used herein
in its broadest sence to mean not only genes associated with classical inherited
diseases, but also those associated with genetic predisposition to disease as well as
infectious or pathogenic states resulting from gene expression by infectious agents or
the effect on host cell gene expression by the presence of such a pathogen or its
30   products Locating disease-associated genes will permit the development of
diagnostic and prognostic reagents and methods, as well as possible therapeutic
regimens, and the discovery of new drugs for treating or preventing the occurrence of
such diseases.

           Methods have been described for the identification of certain novel
35   gene sequences, referred to as Expressed Sequence Tags (EST) [see, e.g., Adams et
al, Science, 252:1651-1656 (1991); and International Patent Application No.
WO93/00353, published January 7, 1993]. Conventially, an EST is a specific cDNA
polynucleotide sequence, or tag, about 150 to 400 nucleotides in length, derived from

a messenger RNA molecule by reverse transcription, which is a marker for, and component of, a human gene actually transcribed *in vivo*. However, as used herein an EST also refers to a genomic DNA fragment derived from an organism, such as a microorganism,the DNA of which lacks intron regions.

5   A variety of techniques have been described for identifying particular gene sequences on the basis of their gene products. For example, several techniques are described in the art [see, e.g., International Patent Application No. WO91/07087, published May 30, 1991]. Additionally, known methods exist for the amplification of desired sequences [see, e.g., International Patent Application No. WO91/17271, 10   published November 14, 1991, among others].

However, at present, there exist no established methods for filling the need in the art for methods and reagents which employ fragments of differentially expressed genes of known, unknown (or previously unrecognized) function or consequence to provide diagnostic and therapeutic methods and reagents for diagnosis 15   and treatment of disease or infection, which conditions are characterized by such genes and gene products. It should be appreciated that it is the expression differences that are diagnostic of the altered state (e.g., predisease, disease, pathogenic, progression or infectious). Such genes associated with the altered state are likely to be the targets of drug discovery, whether the genes are the cause or the effect of the 20   condition, identification of such genes provides insight into which·gene expression needs to be re-altered in order to reestablished the healthy state.

Summary of the Invention

In one aspect, the invention provides methods for identifying gene(s) 25   which are differentially expressed, for example, in a normal healthy organism and an organism having a disease. The method involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences obtained from either analogous cells, tissues or organs of a healthy organism and a diseased organism and a defined set of 30   oligonucleotide/polynucleotide/polynucleotide sequence probes from either an healthy organism or a diseased organism immobilized on a support. Those defined oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined the collection of partial cDNA sequences (ESTs). The differences between the hybridization 35   patterns permit identification of those particular EST or gene-specific oligonucleotide/polynucleotide sequences associated with differential expression, and the identification of the EST permits identification of the clone from which it was

2

derived and using ordinary skill further cloning and, if desired, sequencing of the full-length cDNA and genomic counterpart, i.e., gene, from which it was obtained.

In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those gene(s) of a pathogen which are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism, hybridized to an oligonucleotide/polynucleotide set representative of the gene coding complement of the pathogen of interest.

In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those ESTs-specific oligonucleotide/polynucleotide sequences of host gene(s) which represent genes being differentially expressed/ altered in expression by the disease state, or infection and are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism of interest.

In a further aspect, the methods described above and in detail below, also provide methods for diagnosis of diseases or infections characterized by differentially expressed genes, the expression of which has been altered as a result of infection by the pathogen or disease causing agent in question. All identified differences provide the basis for diagnostic testing be it the altered expression of endogenous genes or the patterned expression of the genes of the infecting organism. Such patterns of altered expression are defined by comparing RNA/cDNA from the two states hybridized against a panel of oligonucleotide/polynucleotides representing the expressed gene component of a cell, tissue, organ or organism as defined by its collection of ESTs.

Yet a further aspect of this invention provides a composition suitable for use in hybridization, which comprises a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence comprising a fragment of an EST isolated from a cDNA or DNA library prepared from at least one selected tissue or cell sample of a healthy (i.e., pre-disease state) animal, at least one analogous sample of an animal having a disease, at least one analogous sample of an animal infected with a pathogen or the pathogen itself, or any combination or multiple combinations thereof.

An additional aspect of the invention provides an isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, and is identified by the methods above. Similarly, an isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal can be identified by the methods above.

3

Yet another aspect of the invention is that it provides not only a means for a static diagnostic but also provides a means for a carrying out the procedure over time to measure disease progression as well as monitoring the efficacy of disease treatment regimes including an toxicological effects thereof.

5   Another aspect of the invention is an isolated protein produced by expression of the gene sequences identified above. Such proteins are useful in therapeutic compositions or diagnostic compositions, or as targets for drug development.

Other aspects and advantages of the present invention are described
10   further in the following detailed description of the preferred embodiments thereof.

Detailed Description of the Invention

The present invention meets the unfulfilled needs in the art by providing methods for the identification and use of gene fragments and genes, even
15   those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. Employing the methods of this invention permits the resulting identification and isolation of such genes by using their corresponding ESTs
20   and thereby also permits the production of protein products encoded by such genes. The genes themselves and/or protein products, if desired, may be employed in the diagnosis or therapy of the disease or infection with which the genes are associated and in the development of new drugs therefor.

It has been appreciated that one or more differentially identified EST
25   or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the ESTs identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. The differences permit the identification of
30   gene products altered in their expression by the disease and represent those products most likely to be targets of therapeutic intervention. Similarly, the product may be of the infecting organism itself and also be an effective target of intervention.

I.   *Definitions.*

35   Several words and phrases used throughout this specification are defined as follows:

As used herein, the term "gene" refers to the genomic nucleotide sequence from which a cDNA sequence is derived, which cDNA produces an EST, as

described below. The term gene classically refers to the genomic sequence, which, upon processing, can produce different cDNAs, e.g., by splicing events. However, for ease of reading, any full-length counterpart cDNA sequence which gives rise to an EST will also be referred to by shorthand herein as a 'gene'.

5   The term "organism" includes without limitation, microbes, plants and animals.

The term "animal" is used in its broadest sense to include all members of the animal kingdom, including humans. It should be understood, however, that according to this invention the same species of animal which provides the biological

10   sample also is the source of the defined immobilized oligonucleotide/polynucleotides as defined below.

The term "pathogen" is defined herein as any molecule or organism which is capable of infecting an animal or plant and replicating its nucleic acid sequences in the cells or tissues of that animal or plant . Such a pathogen is generally

15   associated with a disease condition in the infected animal or plant. Such pathogens may include viruses, which replicate intra- or extra-cellularly, or other organisms, such as bacteria, fungi or parasites, which generally infect tissues or the blood. Certain pathogens or microorganisms are known to exist in sequential and distinguishable stages of development, e.g., latent stages, infective stages, and stages

20   which cause symptomatic diseases. In these different stages, the pathogens are anticipated to express differentially certain genes and/or turn on or off host cell gene expression.

As used herein, the term "disease" or "disease state" refers to any condition which deviates from a normal or standardized healthy state in an organism

25   of the same species in terms of differential expression of the organism's genes. In other words, a disease state can be any illness or disorder be it of genetic or environmental origin , for example, an inherited disorder such as certain breast cancers, or a disorder which is characterized by expression of gene(s) normally in an inactive, 'turned off' state in a healthy animal, or a disorder which is characterized by

30   under-expression or no expression of gene(s) which is normally activated or 'turned on' in a normal healthy animal. Such differential expression of genes may also be detected in a condition caused by infection, inflammation, or allergy, a condition caused by development or aging of the animal, a condition caused by administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects

35   gene expression. Essentially, the methods described herein can be adapted to detect differential gene expression resulting from any cause, by manipulation of the defined oligonucleotide/polynucleotides and the samples tested as described below. The

concept of disease or disease state also includes its temporal aspects in terms of progression and treatment.

The phrase "differentially expressed" refers to those situations in which a gene transcript is found in differing numbers of copies, or in activated vs

5   inactivated states, in different cell types or tissue types of an organism, having a selected disease as contrasted to the levels of the gene transcript found in the same cells or tissues of a healthy organism. Genes may be differentially expressed in differing states of activation in microorganisms or pathogens in different stages of development. For example, multiple copies of gene transcripts may be found in an

10 organism having a selected disease, while only one, or significantly fewer copies, of the same gene transcript are found in a healthy organism, or vice-versa.

As used herein, the term "solid support" refers to any known substrate which is useful for the immobilization of large numbers of oligonucleotide/polynucleotide sequences by any available method to enable

15 detectable hybridization of the immobilized oligonucleotide/polynucleotide sequences with other polynucleotide sequences in a sample. Among a number of available solid supports, one desirable example is the supports described in International Patent Application No. WO91/07087, published May 30, 1991.Also useful are suports such as but not limited to nitrocellulose, mylein, glass, silica ans Pall Biodyne C® It is

20 also anticipated that improvements yet to be made to conventional solid supports may also be employed in this invention.

The term "surface" means any generally two-dimensional structure on a solid support to which the desired oligonucleotide/polynucleotide sequence is attached or immobilized. A surface may have steps, ridges, kinks, terraces and the

25 like.

As used herein, the term "predefined region" refers to a localized area on a surface of a solid support on which is immobilized one or multiple copies of a particular oligonucleotide/polynucleotide sequence and which enables the identification of the oligonucleotide/polynucleotide at the position, if hybridization of

30 that oligonucleotide/polynucleotide to a sample polynucleotide occurs.

By "immobilized" refers to the attachment of the oligonucleotide/polynucleotide to the solid support. Means of immobilization are known and conventional to those of skill in the art, and may depend on the type of support being used.

35 By "EST" or "Expressed Sequence Tag" is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides of a longer sequence obtained from a genomic or cDNA library prepared from a selected cell, cell type, tissue or tissue type, organ or organism which longer

6

sequence corresponds to an mRNA of a gene found in that library. An EST is generally DNA. One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000-100,000 in an 5 animal such as a human. Further background and information on the construction of ESTs is described in M. D. Adams et al, Science, 252:1651-1656 (1991); and International Application Number PCT/US92/05222 (January 7, 1993).

As used herein, the term "defined oligonucleotide/polynucleotide sequence" refers to a known nucleotide sequence fragment of a selected EST or gene. 10 This term is used interchangeably with the term "fragments of EST". These sequential sequences are generally comprised of between about 15 to about 45 nucleotides and more preferably between about 20 to about 25 nucleotides in length. Thus any single EST of 300 nucleotides in length may provide about 280 different defined oligonucleotide/polynucleotide sequences of 20 nucleotides in length (e.g., 15 20-mers). The lengths of the defined oligonucleotide/polynucleotides may be readily increased or decreased as desired or needed, depending on the limitations of the solid support on which they may be immobilized or the requirements of the hybridization conditions to be employed.The length is generally guided by the principle that it should be of sufficient length to insure that it is one average only represented once in 20 the population to be examined. Generally, these defined oligonucleotide/polynucleotides are RNA or DNA and are preferably derived from the anti-sense strand of the EST sequence or from a corresponding mRNA sequence to enable their hybridization with samples of RNA or DNA. Modified nucleotides may be incorporated to increase stability and hybridization properties.

25 By the term "plurality of defined oligonucleotide/polynucleotide sequences" is meant the following. A surface of a solid support may immobilize a large number of "defined oligonucleotide/polynucleotides". For example, depending upon the nature of the surface, it can immobilize from about 300 to upwards of 60,000 defined 20-mer oligonucleotide/polynucleotides. It is anticipated that future 30 improvements to solid surfaces will permit considerably larger such pluralities to be immobilized on a single surface. A "plurality" of sequences refers to the use on any one solid support of multiple different defined oligonucleotide/polynucleotides from a single EST from a selected library, as well as multiple different defined oligonucleotide/polynucleotides from different ESTs from the same library or many 35 libraries from the same or different tissues, and may also include multiple identical copies of defined oligonucleotide/polynucleotides. Ultimately a pluarality has at least one oligonucleotide/polynucleotide per expressed gene in the entire organism For example, from a library producing about 5,000-10,000 ESTs, a single support can

7

include at least about 1-20 defined oligonucleotide/polynucleotides representing every EST in that library. The composition of defined oligonucleotide/polynucleotides which make up a surface according to this invention may be selected or designed as desired.

5   The term "sample" is employed in the description of this invention in several important ways. As used herein, the term "sample" encompasses any cell or tissue from an organism. Any desired cell or tissue type in any desired state may be selected to form a sample. For example, the sample cell desired may be a human T cell; the desired cell type for use in this invention may be a quiescent T cell or an
10   activated T cell.

By the phrase "analogous sample" or "analogous cell or tissue" is meant that according to this invention when the ESTs which provide the defined oligonucleotide/polynucleotides are produced from a cDNA library prepared from a single tissue or cell type source sample, e.g., liver tissue of a human, then the samples
15   used to hybridize to those immobilized defined oligonucleotide/polynucleotides are preferably provided by the same type of sample from either a healthy or diseased animal, i.e., liver tissue of a healthy human and liver tissue of a diseased or infected human or from a human suspected of having that disease or infection. Alternatively, if the surface contains defined oligonucleotide/polynucleotides from multiple cells or
20   tissues, then the "samples" which are hybridized thereto can be but are not limited to samples obtained from analogous multiple tissues or cells.

By the term "detectably hybridizing" means that the sample from the healthy organism or diseased or infected organism is contacted with the defined oligonucleotide/polynucleotides on the surface for sufficient time to permit the
25   formation of patterns of hybridization on the surfaces caused by hybridization between certain polynucleotide sequences in the samples with the certain immobilized defined oligonucleotide/polynucleotides. These patterns are made detectable by the use of available conventional techniques, such as fluorescent labelling of the samples. Preferably hybridization takes place under stringent conditions, e.g., revealing
30   homologies of about 95%. However, if desired, other less stringent conditions may be selected. Techniques and conditions for hybridization at selected stringencies are well known in the art [see, e.g., Sambrook et al, Molecular Cloning. A Laboratory Manual., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989)].

35   *II.   Compositions of The Invention*

The present invention is based upon the use of ESTs from any desired cell or tissue in known technologies for oligonucleotide/polynucleotide hybridization.

A.   ESTs

An EST, as defined above, is for an animal, a sequence from a cDNA clone that corresponds to an mRNA. The EST sequences useful in the present invention are isolated preferably from cDNA libraries using a rapid screening and sequencing technique. Custom made cDNA libraries are made using known techniques. See, generally, Sambrook et al, cited above. Briefly, mRNA from a selected cell or tissue is reverse transcribed into complementary DNA (cDNA) using the reverse transcriptase enzyme and made double-stranded using RNase H coupled with DNA polymerase or reverse transcriptase. Restriction enzyme sites are added to the cDNA and it is cloned into a vector. The result is a cDNA library. Alternatively, commercially available cDNA libraries may be used. Libraries of cDNA can also be generated from recombinant expression of genomic DNA using known techniques, including polymerase chain reaction-derived techniques.

ESTs (which can range from about 150 to about 500 nucleotides in length, preferably about 300 nucleotides) can be obtained through sequence analysis from either end of the cDNA insert. Desirably, the DNA libraries used to obtain ESTs use directional cloning methods so that either the 5' end of the cDNA (likely to contain coding sequence) or the 3' end (likely to be a non-coding sequence) can be selectively obtained.

In general, the method for obtaining ESTs comprises applying conventional automated DNA sequencing technology to screen clones, advantageously randomly selected clones, from a cDNA library. The cDNA libraries from the desired tissue can be preprocessed, or edited, by conventional techniques to reduce repeated sequencing of high and intermediate abundance clones and to maximize the chances of finding rare messages from specific cell populations. Preferably, preprocessing includes the use of defined composition prescreening probes, e.g., cDNA corresponding to mitochondria, abundant sequences, ribosomes, actins, myelin basic polypeptides, or any other known high abundance peptide. These prescreening probes used for preprocessing are generally derived from known ESTs. Other useful preprocessing techniques include subtraction hybridization, which preferentially reduces the population of highly represented sequences in the library [e.g., see Fargnoli et al, Anal. Biochem., 187:364 (1990)] and normalization, which results in all sequences being represented in approximately equal proportions in the library [Patanjali et al, Proc. Natl. Acad. Sci. USA, 88:1943 (1991)]. Additional prescreening/differential screening approaches are known to those skilled in the art.

ESTs can then be generated from partial DNA sequencing of the selected clones. The ESTs useful in the present invention are preferably generated using low redundancy of sequencing, typically a single sequencing reaction. While

single sequencing reactions may have an accuracy as low as 90%, this nevertheless provides sufficient fidelity for identification of the sequence and design of PCR primers.

If desired, the location of an EST in a full length cDNA is determined by analyzing the EST for the presence of coding sequence. A conventional computer program is used to predict the extent and orientation of the coding region of a sequence (using all six reading frames). Based on this information, it is possible to infer the presence of start or stop codons within a sequence and whether the sequence is completely coding or completely non-coding or a combination of the two. If start or stop codons are present, then the EST can cover both part of the 5'-untranslated or 3'-untranslated part of the mRNA (respectively) as well as part of the coding sequence. If no coding sequence is present, it is likely that the EST is derived from the 3' untranslated sequence due to its longer length and the fact that most cDNA library construction methods are biased toward the 3' end of the mRNA. It should be understood that both coding and non-coding regions may provide ESTs equally useful in the described invention.

A number of specific ESTs suitable for use in the present invention are described above Adams et al (supra), which may be incorporated by reference herein, to describe non-essential examples of desirable ESTs. Other ESTs exist in the art which may also be useful in this invention, as will ESTs yet to be developed by these known techniques.

B.   *Preparing the Solid Support of the Invention*

Oligonucleotide sequences which are fragments of defined sequence are derived from each EST by conventional means, e.g., conventional chemical synthesis or recombinant techniques. Each defined oligonucleotide/polynucleotide sequence as described above is a fragment, can be, but is not necessarily an anti-sense fragment, of an EST isolated from a DNA library prepared from a selected cell or tissue type from a selected animal. For use in the present invention, it is presently preferred that the defined oligonucleotide/polynucleotide sequences are 20-25mers. As described above, for each EST a number of such 20-25mers may be generated. The lengths may vary as described above as well as the composition. For example oligonucleotide/polynucleotides can be modified based on the Oligo 4.0 or simiolar programs to predict hybridization potential or to include modifieid nucleotides for the reasons given above. It is alos appreciated that large DNA segments may be employed including entire ESTs or even full length genes particular when inserted into cloning vectors.

10

A plurality of these defined oligonucleotide/polynucleotide sequences are then attached to a selected solid support conventionally used for the attachment of nucleotide sequences again by known means. In contrast to other technologies available in the art, this support is designed to contain defined, not

5   random, oligonucleotide/polynucleotide sequences. The EST fragments, or defined oligonucleotide/polynucleotide sequences, immobilized on the solid support can include fragments of one or more ESTs from a library of at least one selected tissue or cell sample of a healthy animal, at least one analogous sample of the animal having a disease, at least one analogous sample of the animal infected with a pathogen, and

10   any combination thereof.

Numerous conventional methods are employed for attaching biological molecules such as oligonucleotide/polynucleotide sequences to surfaces of a variety of solid supports. See, e.g., Affinity Techniques, Enzyme Purification: Part B. Methods in Enzymology, Vol. 34, ed. W.B. Jakoby, M. Wilcheck, Acad. Press,

15   NY (1974); Immobilized Biochemicals and Affinity Chromatography, Advances in Experimental Medicine and Biology, vol. 42, ed. R. Dunlap, Plenum Press, NY (1974); U. S. Patent No. 4,762,881; U. S. Patent No. 4,542,102; European Patent Publication No. 391,608 (October 10, 1990); U. S. Patent No. 4,992,127 (Nov. 21, 1989).

20   One   desirable   method   for   attaching oligonucleotide/polynucleotide sequences derived from ESTs to a solid support is described in International Application No. PCT/US90/06607 (published May 30, 1991). Briefly, this method involves forming predefined regions on a surface of a solidsupport, where the predefined regions are capable of immobilizing ESTs. The

25   methods make use of binding substances attached to the surface which enable selective activation of the predefined regions. Upon activation, these binding substances   become   capable   of   binding   and   immobilizing oligonucleotide/polynucleotides based on EST or longer gene sequences.

Any of the known solid substrates suitable for binding

30   oligonucleotide/polynucleotides at pre-defined regions on the surface thereof for hybridization and methods for attaching the oligonucleotide/polynucleotides thereto may be employed by one of skill in the art according to this invention. Similarly, known conventional methods for making hybridization of the immobilized oligonucleotide/polynucleotides detectable, e.g., fluorescence, radioactivity,

35   photoactivation, biotinylation, solid state circuitry, and the like may be used in this invention.

Thus, by resorting to known techniques, the invention provides a composition suitable for use in hybridization which consists of a surface of a solid

11

support on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. For example, one composition of this invention is a solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type, e.g., a human

5   stem cell, or a single tissue, e.g., human liver, from a healthy human. Still another composition of this invention is another solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type or a tissue from a human having a selected disease or predispositon to a selected disease, e.g., liver cancer.

10   Another embodiment of the compositions of this invention include a single solid support having oligonucleotides of ESTs from both single cell or single tissue libraries from both a healthy and diseased human. Still other embodiments include a single support on which are immobilized oligos of EST fragments from more than one tissue or cell library from a healthy human or a single

15   support on which are immobilized more than one tissue or cell library from both healthy and diseased animals or humans. A preferred composition of this invention is anticipated to be a single support containing oligos of ESTs for all known cells and tissues from a selected organism.

20   *III.   The Methods of the Invention*
              A.   Identification of Genes

                       The present invention employs the compositions described above in methods for identifying genes which are differentially expressed in a normal healthy organism and an organism having a disease or infection. These methods may

25   be employed to detect such genes, regardless of the state of knowledge about the function of the gene. The method of this invention by use of the compositions containing multiple defined EST fragments from a single gene as described above is able to detect levels of expression of genes or in other cases simply the expression or lack thereof, which differ between normal, healthy organisms and organisms having a

30   selected disease, disorder or infection.

                       One such method employs a first surface of a solid support on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences, described above, of EST or longer gene fragment isolated from a cDNA library prepared from at least one selected tissue or

35   cell sample of a healthy animal (the "healthy test surface") and a second such surface on which is immobilized at pre-defined regions a plurality of defined oligonucleotide/polynucleotide sequences of EST or longer gene fragment isolated from at least one analogous tissue of an animal having a selected disease (the "disease

test surface"). These test surfaces may be standardized for the selected animal or selected cell or tissue sample from that animal (i.e., they are prescreened for polymorphisms in the species population).

Polynucleotide sequences are then isolated from mRNA and/or
5   cDNA from a biological sample from a known healthy animal ("healthy control") and a second sample is similarly prepared from a sample from a known diseased animal ("disease sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides.

According to the method the healthy control sample is
10   contacted with one set of the healthy test surface and the disease test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed between the nucleotides of healthy control and the healthy test surface and a second
15   hybridization pattern formed between the nucleotides of healthy control sample and the disease test surface.

In a similar manner, the disease sample is detectably hybridized to another set of healthy test and disease test surfaces, forming a third hybridization pattern between the disease sample and healthy test surface and a fourth hybridization
20   pattern between the disease sample and the disease test surface.

Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the   hybridization   patterns   at   pre-defined   regions.   The
25   oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTor longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

In another embodiment of the method of this invention, the same process is employed, with the exception that plurality of defined
30   oligonucleotide/polynucleotide sequences forming the healthy test sample and the disease test sample surfaces are immobilized on a single solid support. For example, each fragment of an EST or longer gene fragment on the surface is isolated from at least two cDNA libraries prepared from a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having a disease.

35   According to this embodiment, the healthy control sample is detectably hybridized to a copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal. Similarly, the disease sample is detectably hybridized to a second

13

copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal.

Comparing the two hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding EST or longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

The identification of one or more ESTs as the source of the defined oligonucleotide/polynucleotide which produced a "difference" in hybridization patterns according to these methods permits ready identification of the gene from which those ESTs were derived. Because oligonuleotides are of sufficient length that they will hybridize under stringent conditions only with a RNA/cDNA for that gene to which they correspond, the oligo can be used to identify the EST and in turn the clone from which it was derived and by subsequent cloning, obtain the sequence of the full-length cDNA and its genomic counterparts, i.e., the gene, from which it was obtained.

In other words, the ESTs identified by the method of this invention can be employed to determine the complete sequence of the mRNA, in the form of transcribed cDNA, by using the EST as a probe to identify a cDNA clone corresponding to a full-length transcript, followed by sequencing of that clone. The EST or the full length cDNA clone can also be used as a probe to identify a genomic clone or clones that contain the complete gene including regulatory and promoter regions, exons, and introns.

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained, rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used a form diagnostic patterns or to identifiy which particular EST is detected. For example, all known ESTs from an organism are used to produce a "master" solid support to which control sample and disease samples are alternately hybridized. One then detects a pattern of hybridization associated with the particular disaease state which then forms the basis of a diagnostic test or the isolation of disease specific ESTs from which the intact gene may be cloned and sequenced leading uiltimately to a defined therapuetic target.

Methods for obtaining complete gene sequences from ESTs are well-known to those of skill in the art. See, generally, Sambrook et al, cited above. Briefly, one suitable method involves purifying the DNA from the clone that was

14

sequ nced to give the EST and labeling the isolated insert DNA. Suitable labeling systems are well known to those of skill in the art [see, eg. Basic Methods in Molecular Biology, L. G. Davis et al, ed., Elsevier Press, NY (1986)]. The labeled EST insert is then used as a probe to screen a lambda phage cDNA library or a

5   plasmid cDNA library, identifying colonies containing clones related to the probe cDNA which can be purified by known methods. The ends of the newly purified clones are then sequenced to identify full length sequences and complete sequencing of full length clones is performed by enzymatic digestion or primer walking. A similar screening and clone selection approach can be applied to clones from a

10   genomic DNA library.

Additionally, an EST or gene identified by this method as associated with inherited disorders can be used to determine at what stage during embryonic development the selected gene from which it is derived is developed by screening embryonic DNA libraries from various stages of development, e.g. 2-cell,

15   8-cell, etc., for the selected gene. As has been mentioned above, the invention may be applied in addtional temporal modes for monitoring the progression of a disease state, the efficacy of a particular treatment modality or the aging process of an individual.

Thus, the methods of this invention permit the identification,

20   isolation and sequencing of a gene which is differentially expressed in a selected disease/infection. As described in more detail below, the identified gene may then be employed to obtain any protein encoded thereby, or may be employed as a target for diagnostic methods or therapeutic approaches to the treatment of the disease, including, e.g., drug development.

25   The same methods as described above for the identification of genes, including genes of unknown function, which are differentially expressed in a disease state, may also be employed to identify other genes of interest. For example, another embodiment of this invention includes a method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with that

30   pathogen or the gene of the host which is altered in its expression as a result of the infection.

One such method employs a healthy test surface as described above, employing defined oligonucleotide/polynucleotides from a sample of a healthy, uninfected animal. The second such surface has immobilized at pre-defined

35   regions thereon a plurality of defined oligonucleotide/polynucleotide sequences of ESTs isolated from at least one analogous tissue or cell sample of an infected animal (the "infection test surface"). Polynucleotide sequences are isolated from a biological sample from a healthy animal ("healthy control") and a second sample is similarly

prepared from an animal infected with the selected pathogen ("infection sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides. It would also be possible to provide samples from the nucleic acid of the pathogen itself.

5   According to the method the healthy control sample is contacted with one set of the healthy test surface and the infection test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed
10   between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the infection test surface.

            In a similar manner, the infection sample is detectably hybridized to another set of healthy test and infection test surfaces, forming a third
15   hybridization pattern between the infection sample and healthy test surface and a fourth hybridization pattern between the infection sample and the infection test surface.

            Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed
20   between the healthy animal and the animal infected with the pathogen by the presence of differences in the hybridization patterns at pre-defined regions. As mentioned differential expression is not required and simple qualitative analysis is possible by reference to gene expression which is simply present or absent.

            A second embodiment of this method parallels the second
25   embodiment of the method as applied to disease above, i.e., the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample surface and the infection test sample surface are immobilized on a single solid support. The resulting first hybridization pattern (healthy control sample with healthy/infection test sample) and second
30   hybridization pattern (infection sample with healthy/infection test sample) permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the infection sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern
35   differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained.

            As described above for the methods for identifying differential gene   expression   between   diseased   and   healthy   animals,   the

16

oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotide sequences are obtained and the genes expressed by the pathogen identified for similar purposes. Other embodiments of these methods may

5   be developed with resort to the teaching herein, by altering the samples which provide the defined oligonucleotide/polynucleotides. For example, an EST, identified with a differentially expressed gene by the method of this invention is also useful in detecting genes expressed in the various stages of an pathogen's development, particularly the infective stage and following the cours of drug treatment and

10   emergence of resistant variants. For example, employing the techniques described above, the EST can be used for detecting a gene in various stages of the parasitic *Plasmodium* species life cycle, which include blood stages, liver stages, and gametocyte stages.

       B.   Diagnostic Methods

15   In addition to use of the methods and compositions of this invention for identifying differentially expressed genes, another embodiment of this invention provides diagnostic methods for diagnosing a selected disease state, or a selected state resulting from aging, exposure to drugs or infection in an animal. According to this aspect of the invention, a first surface, described as the healthy test

20   surface above, and a second surface, described as the disease test surface or infection test surface, are prepared depending on the disease or infection to be diagnosed. The same processes of detectable hybridization to a first and second set of these surfaces with the healthy control sample and disease/infection sample are followed to provide the four above-described hybridization patterns, i.e., healthy control sample with

25   healthy test surface; healthy control sample with disease/infection test surface; disease/infection sample with healthy test surface; and disease/infection sample with disease/infection test surface.

                      The diagnosis of disease or infection is provided by comparing the four hybridization patterns. Substantial differences between the first and third

30   hybridization patterns, respectively, and the second and fourth hybridization patterns, respectively, indicate the presence of the selected disease or infection in said animal. Substantial similarities in the first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

                      A similar embodiment utilizes the single surface bearing both

35   the healthy test surface defined oligonucleotide/polynucleotides and the disease/infection test surface defined oligonucleotide/polynucleotides as described above. Parallel process steps as described above for detection of genes differentially expressed in disease and infected states are followed, resulting in a first hybridization

17

pattern (healthy control sample with single healthy and disease/infection test sample) and a second hybridization pattern (disease/infection sample with another copy of the single healthy and disease/infection test sample).

Diagnosis is accomplished by comparing the two hybridization

5   patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This like many of the foregoing embodiments may use known or unknown ESTs derived from many libraries.

10   C.   Other Methods of the Invention

As is obvious to one of skill in the art upon reading this disclosure, the compositions and methods of this invention may also be used for other similar purposes. For example, the general methods and compositions may be adapted easily by manipulation of the samples selected to provide the standardized

15   defined oligonucleotide/polynucleotides, and selection of the samples selected for hybridization thereto. One such modification is the use of this invention to identify cell markers of any type, e.g., markers of cancer cells, stem cell markers, and the like. Another modification involves the use of the method and compositions to generate hybridization patterns useful for forensic identification or an 'expression fingerprint'

20   of genes for identification of one member of a species from another. Similarly, the methods of this invention may be adapted for use in tissue matching for transplantation purposes as well as for molecular histology, i.e., to enable diagnosis of disease or disorders in pathology tissue samples such as biopsies. Still another use of this method is in monitoring the effects of development and aging upon the gene

25   expression   in   a   selected   animal,   by   preparing   surfaces   bearing oligonucleotide/polynucleotides prepared from samples of standardized younger members of the species being tested. Additionally the patient can serve as an internal control by virtue of having the method applied to blood samples every 5-10 years during his lifetime.

30   Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal, especially humans. Because the method can be readily adapted by altering the above parameters, it can essentially be employed to identify differentially expressed genes of any organism, at any stage of development, and

35   under the influence of any factor which can affect gene expression.

IV.   The Genes and Proteins Identified

Application of the compositions and methods of this invention as above described also provide other compositions, such as any isolated gene sequence which is differentially expressed between a normal healthy animal and an animal having a disease or infection. Another embodiment of this invention is any isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal. Similarly an embodiment of this invention is any gene sequence identified by the methods described herein.

These gene sequences may be employed in conventional methods to produce isolated proteins encoded thereby. To produce a protein of this invention, the DNA sequences of a desired gene identified by the use of the methods of this invention or portions thereof are inserted into a suitable expression system. Desirably, a recombinant molecule or vector is constructed in which the polynucleotide sequence encoding the protein is operably linked to a heterologous expression control sequence permitting expression of the human protein. Numerous types of appropriate expression vectors and host cell systems are known in the art for mammalian (including human) expression, insect, e.g., baculovirus expression, yeast, fungal, and bacterial expression, by standard molecular biology techniques.

The transfection of these vectors into appropriate host cells, whether mammalian, bacterial, fungal, or insect, or into appropriate viruses, can result in expression of the selected proteins. Suitable host cells or cell lines for transfection, and viruses, as well as methods for the construction and transfection of such host cells and viruses are well-known. Suitable methods for transfection, culture, amplification, screening, and product production and purification are also known in the art.

The genes and proteins identified by this invention can be employed, if desired in diagnostic compositions useful for the diagnosis of a disease or infection using conventional diagnostic assays. For example, a diagnostic reagent can be developed which detectably targets a gene sequence or protein of this invention in a biological sample of an animal. Such a reagent may be a complementary nucleotide sequence, an antibody (monoclonal, recombinant or polyclonal), or a chemically derived agonist or antagonist. Alternatively, the proteins and polynucleotide sequences of this invention, fragments of same, or complementary sequences thereto, may themselves be useful as diagnostic reagents for diagnosing disease states with which the ESTs of the invention are associated. These reagents may optionally be labelled using diagnostic labels, such as radioactive labels, colorimetric enzyme label systems and the like conventionally used in diagnostic or therapeutic methods, e.g, Northern and Western blotting, antigen-antibody binding and the like. The selection of the appropriate assay format and label system is within the skill of the art and may

readily be chosen without requiring additional explanation by resort to the wealth of art in the diagnostic area.

Additionally, genes and proteins identified according to this invention may be used therapeutically. For example, the EST-containing gene sequences may be useful in gene therapy, to provide a gene sequence which in a disease is not properly or sufficiently expressed. In such a method, a selected gene sequence of this invention is introduced into a suitable vector or other delivery system for delivery to a cell containing a defect in the selected gene. Suitable delivery systems are well known to those of skill in the art and enable the desired EST or gene to be incorporated into the target cell and to be translated by the cell. The EST or gene sequence may be introduced to mutate the existing gene by recombination or provide an active copy thereof in addition to the inactive gene to replace its function.

Alternatively, a protein encoded by an EST or gene of the invention may be useful as a therapeutic reagent for delivery of a biologically active protein, particularly when the disease state is associated with a deficiency of this protein. Such a protein may be incorporated into an appropriate therapeutic formulation, alone or in combination with other active ingredients. Methods of formulating such therapeutic compositions, as well as suitable pharmaceutical carriers, and the like, are well known to those of skill in the art. Still an additional method of delivering the missing protein encoded by an EST, or the gene from which a selected EST was derived, involves expressing it directly *in vivo*. Systems for such *in vivo* expression are well known in the art.

Yet another use of the ESTs, genes identified according to the methods of this invention, or the proteins encoded thereby is a target for the screening and development of natural or synthetic chemical compounds which have utility as therapeutic drugs for the treatment of disease states associated with the identified genes and ESTs derived therefrom. As one example, a compound capable of binding to such a protein encoded by such a gene and either preventing or enhancing its biological activity may be a useful drug component for the treatment or prevention of such disease states.

Conventional assays and techniques may be used for the screening and development of such drugs. As one example, a method for identifying compounds which specifically bind to or inhibit or activate proteins encoded by these gene sequences can include simply the steps of contacting a selected protein or gene product, with a test compound to permit binding of the test compound to the protein; and determining the amount of test compound, if any, which is bound to the protein. Such a method may involve the incubation of the test compound and the protein immobilized on a solid support. Still other conventional methods of drug screening

20

can involve employing a suitable computer program to determine compounds having similar or complementary chemical structures to that of the gene product or portions thereof and screening those compounds either for competitive binding to the protein to detect enhanced or decreased activity in the presence of the selected compound.

5   Thus, through use of such methods, the present invention is anticipated to provide compounds capable of interacting with these genes, ESTs, or encoded proteins, or fragments thereof, and either enhancing or decreasing the biological activity, as desired. Such compounds are believed to be encompassed by this invention.

10   Numerous modifications and variations of the present invention are included in the above-identified specification and are expected to be obvious to one of skill in the art. Such modifications and alterations to the compositions and processes of the present invention are believed to be encompassed in the scope of the claims appended hereto.

15

WHAT IS CLAIMED IS:

1.   A method for identifying genes which are differentially expressed in two different pre-determined states of an organism comprising:

5   a. providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a first

10   state and present in excess relative to the polynucleotide to be hybridized;

    b. providing a second surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library

15   prepared from at least one selected cell, tissue, organ or organism sample in a second state and present in excess relative to the polynucleotide to be hybridized;

    c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a said organism in said first state, said sample selected from sources analogous to the sources of step (a), said

20   hybridization sufficient to form a first and second hybridization pattern on each said first and second surface,

    d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from said organism in said second state, said sample selected from sources analogous to the sources of step (c), said

25   hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface,

    e.   comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;

30   f.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

35

2. The method according to Claim 1 wherein said first and second states are respectively healthy and disease; pathogen uninfected and pathogen infected; a first progression state and a second progression of a disease or infection; a first treatment state and a second treatment state of a disease or infection; or a first developmental

5   and a second developmental state.

3. The method according to Claim 1 wherein said organism is a plant or an animal.

10   4. The method according to Claim 3 wherein said aniaml is a human.

5.   A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a.   providing a first surface on which is immobilized at pre-
15   defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a healthy animal and present in excess relative to the polynucleotide to be
20   hybridized;

b.   providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated
25   from a DNA library prepared from at least one selected cell, tissue, organ or organism sample from an animal having said disease and present in excess relative to the polynucleotide to be hybridized;

c.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample
30   selected from sources analogous to the sources of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

23

d.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each

5   said first and second surface,

e.   comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f.   identifying the oligonucleotide/polynucleotides on each surface

10   which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

15   6.   A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a.   providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,

20   an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized;

25   b.   detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c.   detectably hybridizing to a second copy of said surface

30   polynucleotide sequences isolated from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d.   comparing the two hybridization patterns, wherein genes differentially expressed in a disease state are identified by the presence of differences

35   in the hybridization patterns at pre-defined regions;

e.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

5

7.   A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a.   providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide

10   sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy, uninfected animal and present in excess relative to the polynucleotide to be hybridized;

15   b.   providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from at least one selected cell, tissue, organ or organism sample of an infected animal;

20   c.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form first and second hybridization patterns on each said first and second surface,

25   d.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form third and fourth hybridization patterns on each said first and second surface,

30   e.   comparing the four hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which

35   the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

8.   A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a.   providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide

5   sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to

10   be hybridized

b.   detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

15   c.   detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d.   comparing the two hybridization patterns, wherein genes of

20   said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

e.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST

25   permits identification of the gene from which the ESTs were derived.

9.   A composition suitable for use in hybridization comprising a solid surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment

30   of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample of a healthy animal, at least one analogous sample of said animal having a disease, at least one analogous sample of said animal infected with a microbial pathogen, and any

35   combination thereof.

26

10.   An isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, identified by the method of claim 1.

5   11.   An isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal identified by the method of claim 7.

12.   A diagnostic composition useful for the diagnosis of a disease comprising a reagent capable of detectably targeting a gene sequence of claim 10 in a 10 biological sample of an animal.

13.   A diagnostic composition useful for the diagnosis of infection by a pathogen comprising a reagent capable of detectably targeting a gene sequence of claim 11 in a biological sample of an animal.
15

14.   An isolated protein produced by expression of a gene sequence of claim 10.

15.   An isolated pathogen protein produced by expression of a gene 20 sequence of claim 11.

16.   A therapeutic composition comprising a protein or fragment thereof selected from the group consisting of a protein of claim 10 and a protein of claim 15.

25   17.   A method for diagnosing a selected disease or infection in an animal comprising:
        a.   providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, 30 an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy animal and present in excess relative to the polynucleotide to be hybridized;
        b.   providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide 35 sequences, each sequence comprising a fragment of an EST isolated from at least one said tissue of an animal having said disease;

27

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second

5   hybridization pattern on each said first and second surface;

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and

10   fourth hybridization pattern on each said first and second surface;

e. comparing the four hybridization patterns, wherein substantial differences between the first and third hybridization patterns and the second and fourth hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and third

15   hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

18.   A method for diagnosing a selected disease or infection in an animal comprising:

20   a.   providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from a DNA library prepared from the group consisting of a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having

25   said disease;

b.   detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

30   c.   detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

35   d.   comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and second hybridization patterns indicates the absence of disease or infection.

International application No.

PCT/US95/01863

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68
US CL :435/6

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, CAS, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ANALYTICAL BIOCHEMISTRY, VOLUME 187, ISSUED 1990, FARGNOLI ET AL, "LOW-RATIO HYBRIDIZATION SUBTRACTION", PAGES 364-373, SEE ENTIRE DOCUMENT.	1-18
Y	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES USA, VOLUME 88, ISSUED MARCH 1991, PATANJALI ET AL, "CONSTRUCTION OF A UNIFORM-ABUNDANCE (NORMALIZED) CDNA LIBRARY", PAGES 1943-1947, SEE ENTIRE DOCUMENT.	1-18
Y	SCIENCE, VOLUME 245, ISSUED 29 SEPTEMBER 1989, OLSON ET AL. "A COMMON LANGUAGE FOR PHYSICAL MAPPING OF THE HUMAN GENOME", PAGES 1434-1435, SEE ENTIRE DOCUMENT.	1-18

| X | Further documents are listed in the continuation of Box C. | | See patent family annex. |

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
03 APRIL 1995	17 MAY 1995

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	EGGERTON CAMPBELL
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCIENCE, VOLUME 252, ISSUED 21 JUNE 1991, ADAMS ET AL, "COMPLEMENTARY DNA SEQUENCING: EXPRESSED SEQUENCE TAGS AND HUMAN GENOME PROJECT", PAGES 1651-1656, SEE ENTIRE DOCUMENT.	1-18

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification [6] : **C12Q 1/68, G06F 15/00**	**A1**	(11) International Publication Number: **WO 95/20681** (43) International Publication Date: 3 August 1995 (03.08.95)

(21) International Application Number: PCT/US95/01160

(22) International Filing Date: 27 January 1995 (27.01.95)

(30) Priority Data:
08/187,530   27 January 1994 (27.01.94)   US
08/282,955   29 July 1994 (29.07.94)   US

(71) Applicant: INCYTE PHARMACEUTICALS, INC. [US/US]; 3330 Hillview Avenue, Palo Alto, CA 94304 (US).

(72) Inventors: SEILHAMER, Jeffrey, J.; 12555 La Cresta, Los Altos Hills, CA 94022 (US). SCOTT, Randal, W.; 13140 Sun-Mor, Mountain View, CA 94040 (US).

(74) Agents: CAGE, Kenneth, L. et al.; Willian Brinks Hofer Gilson & Lione, 2000 K Street, N.W., Suite 200, Washington, DC 20006-1809 (US).

(81) Designated States: AM, AU, BB, BG, BR, BY, CA, CN, CZ, EE, FI, GE, HU, JP, KG, KP, KR, KZ, LK, LR, LT, LV, MD, MG, MN, MX, NO, NZ, PL, RO, RU, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ).

Published
With international search report.

(54) Title: COMPARATIVE GENE TRANSCRIPT ANALYSIS

(57) Abstract

A method and system for quantifying the relative abundance of gene transcripts in a biological specimen. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs (gene transcript imaging analysis). Another embodiment of the method produces a gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, the gene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens.

COMPARATIVE GENE TRANSCRIPT ANALYSIS

1. FIELD OF INVENTION

The present invention is in the field of molecular
biology and computer science; more particularly, the
5   present invention describes methods of analyzing gene
transcripts and diagnosing the genetic expression of cells
and tissue.

2. BACKGROUND OF THE INVENTION

Until very recently, the history of molecular biology
10 has been written one gene at a time. Scientists have
observed the cell's physical changes, isolated mixtures
from the cell or its milieu, purified proteins, sequenced
proteins and therefrom constructed probes to look for the
corresponding gene.
15   Recently, different nations have set up massive
projects to sequence the billions of bases in the human
genome. These projects typically begin with dividing the
genome into large portions of chromosomes and then
determining the sequences of these pieces, which are then
20 analyzed for identity with known proteins or portions
thereof, known as motifs. Unfortunately, the majority of
genomic DNA does not encode proteins and though it is
postulated to have some effect on the cell's ability to
make protein, its relevance to medical applications is not
25 understood at this time.
A third methodology involves sequencing only the
transcripts encoding the cellular machinery actively
involved in making protein, namely the mRNA. The advantage
is that the cell has already edited out all the non-coding
30 DNA, and it is relatively easy to identify the protein-
coding portion of the RNA. The utility of this approach
was not immediately obvious to genomic researchers. In
fact, when cDNA sequencing was initially proposed, the
method was roundly denounced by those committed to genomic
35 sequencing. For example, the head of the U.S. Human Genome
project discounted CDNA sequencing as not valuable and
refused to approve funding of projects.
In this disclosure, we teach methods for analyzing
DNA, including cDNA libraries. Based on our analyses and

research, we see each individual gene product as a "pixel"
of information, which relates to the expression of that,
and only that, gene. We teach herein, methods whereby the
individual "pixels" of gene expression information can be
5   combined into a single gene transcript "image," in which
each of the individual genes can be visualized
simultaneously and allowing relationships between the gene
pixels to be easily visualized and understood.

We further teach a new method which we call electronic
10   subtraction. Electronic subtraction will enable the gene
researcher to turn a single image into a moving picture,
one which describes the temporality or dynamics of gene
expression, at the level of a cell or a whole tissue. It
is that sense of "motion" of cellular machinery on the
15   scale of a cell or organ which constitutes the new
invention herein. This constitutes a new view into the
process of living cell physiology and one which holds great
promise to unveil and discover new therapeutic and
diagnostic approaches in medicine.
20   We teach another method which we call "electronic
northern," which tracks the expression of a single gene
across many types of cells and tissues.

Nucleic acids (DNA and RNA) carry within their
sequence the hereditary information and are therefore the
25   prime molecules of life. Nucleic acids are found in all
living organisms including bacteria, fungi, viruses, plants
and animals. It is of interest to determine the relative
abundance of different discrete nucleic acids in different
cells, tissues and organisms over time under various
30   conditions, treatments and regimes.

All dividing cells in the human body contain the same
set of 23 pairs of chromosomes. It is estimated that these
autosomal and sex chromosomes encode approximately 100,000
genes. The differences among different types of cells are
35   believed to reflect the differential expression of the
100,000 or so genes. Fundamental questions of biology
could be answered by understanding which genes are
transcribed and knowing the relative abundance of
transcripts in different cells.

Previously, the art has only provided for the analysis
of a few known genes at a time by standard molecular
biology techniques such as PCR, northern blot analysis, or
other types of DNA probe analysis such as _in situ_
5   hybridization. Each of these methods allows one to analyze
the transcription of only known genes and/or small numbers
of genes at a time. Nucl. Acids Res. 19, 7097-7104 (1991);
Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18,
2789-92 (1989); European J. Neuroscience 2, 1063-1073
10   (1990); Analytical Biochem. 187, 364-73 (1990); Genet.
Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33
(1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988);
Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci.
USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27
15   (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988);
Nucl. Acids Res. 16, 10937 (1988).

Studies of the number and types of genes whose
transcription is induced or otherwise regulated during cell
processes such as activation, differentiation, aging, viral
20   transformation, morphogenesis, and mitosis have been
pursued for many years, using a variety of methodologies.
One of the earliest methods was to isolate and analyze
levels of the proteins in a cell, tissue, organ system, or
even organisms both before and after the process of
25   interest. One method of analyzing multiple proteins in a
sample is using 2-dimensional gel electrophoresis, wherein
proteins can be, in principle, identified and quantified as
individual bands, and ultimately reduced to a discrete
signal. At present, 2-dimensional analysis only resolves
30   approximately 15% of the proteins. In order to positively
analyze those bands which are resolved, each band must be
excised from the membrane and subjected to protein sequence
analysis using Edman degradation. Unfortunately, most of
the bands were present in quantities too small to obtain a
35   reliable sequence, and many of those bands contained more
than one discrete protein. An additional difficulty is
that many of the proteins were blocked at the
amino-terminus, further complicating the sequencing
process.

Analyzing differentiation at the gene transcription
level has overcome many of these disadvantages and
drawbacks, since the power of recombinant DNA technology
allows amplification of signals containing very small

5   amounts of material. The most common method, called
"hybridization subtraction," involves isolation of mRNA
from the biological specimen before (B) and after (A) the
developmental process of interest, transcribing one set of
mRNA into cDNA, subtracting specimen B from specimen A

10   (mRNA from cDNA) by hybridization, and constructing a cDNA
library from the non-hybridizing mRNA fraction. Many
different groups have used this strategy successfully, and
a variety of procedures have been published and improved
upon using this same basic scheme. Nucl. Acids Res. 19,

15   7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990);
Nucl. Acids Res. 18, 2789-92 (1989); European J.
Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187,
364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990);
GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85,

20   1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc.
Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res.
19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42
(1988); Nucl. Acids Res. 16, 10937 (1988).

Although each of these techniques have particular

25   strengths and weaknesses, there are still some limitations
and undesirable aspects of these methods: First, the time
and effort required to construct such libraries is quite
large. Typically, a trained molecular biologist might
expect construction and characterization of such a library

30   to require 3 to 6 months, depending on the level of skill,
experience, and luck. Second, the resulting subtraction
libraries are typically inferior to the libraries
constructed by standard methodology. A typical
conventional cDNA library should have a clone complexity of

35   at least 10^6 clones, and an average insert size of 1-3 kB.
In contrast, subtracted libraries can have complexities of
10^2 or 10^3 and average insert sizes of 0.2 kB. Therefore,
there can be a significant loss of clone and sequence
information associated with such libraries. Third, this

approach allows the researcher to capture only the genes
induced in specimen A relative to specimen B, not
vice-versa, nor does it easily allow comparison to a third
specimen of interest (C). Fourth, this approach requires

5   very large amounts (hundreds of micrograms) of "driver"
mRNA (specimen B), which significantly limits the number
and type of subtractions that are possible since many
tissues and cells are very difficult to obtain in large
quantities.

10   Fifth, the resolution of the subtraction is dependent
upon the physical properties of DNA:DNA or RNA:DNA
hybridization. The ability of a given sequence to find a
hybridization match is dependent on its unique CoT value.
The CoT value is a function of the number of copies

15   (concentration) of the particular sequence, multiplied by
the time of hybridization. It follows that for sequences
which are abundant, hybridization events will occur very
rapidly (low CoT value), while rare sequences will form
duplexes at very high CoT values. CoT values which allow

20   such rare sequences to form duplexes and therefore be
effectively selected are difficult to achieve in a
convenient time frame. Therefore, hybridization
subtraction is simply not a useful technique with which to
study relative levels of rare mRNA species. Sixth, this

25   problem is further complicated by the fact that duplex
formation is also dependent on the nucleotide base
composition for a given sequence. Those sequences rich in
G + C form stronger duplexes than those with high contents
of A + T. Therefore, the former sequences will tend to be

30   removed selectively by hybridization subtraction. Seventh,
it is possible that hybridization between nonexact matches
can occur. When this happens, the expression of a
homologous gene may "mask" expression of a gene of
interest, artificially skewing the results for that

35   particular gene.
          Matsubara and Okubo proposed using partial cDNA
sequences to establish expression profiles of genes which
could be used in functional analyses of the human genome.
Matsubara and Okubo warned against using random priming, as

it creates multiple unique DNA fragments from individual
mRNAs and may thus skew the analysis of the number of
particular mRNAs per library. They sequenced randomly
selected members from a 3'-directed cDNA library and
5   established the frequency of appearance of the various
ESTs. They proposed comparing lists of ESTs from various
cell types to classify genes. Genes expressed in many
different cell types were labeled housekeepers and those
selectively expressed in certain cells were labeled cell-
10   specific genes, even in the absence of the full sequence of
the gene or the biological activity of the gene product.
      The present invention avoids the drawbacks of the
prior art by providing a method to quantify the relative
abundance of multiple gene transcripts in a given
15   biological specimen by the use of high-throughput
sequence-specific analysis of individual RNAs and/or their
corresponding cDNAs.
      The present invention offers several advantages over
current protein discovery methods which attempt to isolate
20   individual proteins based upon biological effects. The
method of the instant invention provides for detailed
diagnostic comparisons of cell profiles revealing numerous
changes in the expression of individual transcripts.
      The instant invention provides several advantages over
25   current subtraction methods including a more complex
library analysis (10^6 to 10^7 clones as compared to 10^3
clones) which allows identification of low abundance
messages as well as enabling the identification of messages
which either increase or decrease in abundance. These
30   large libraries are very routine to make in contrast to the
libraries of previous methods. In addition, homologues can
easily be distinguished with the method of the instant
invention.
      This method is very convenient because it organizes a
35   large quantity of data into a comprehensible, digestible
format. The most significant differences are highlighted
by electronic subtraction. In depth analyses are made more
convenient.

The present invention provides several advantages over previous methods of electronic analysis of cDNA. The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed.
5 In such a case, new low-frequency transcripts are discovered and tissue typed.

High resolution analysis of gene expression can be used directly as a diagnostic profile or to identify disease-specific genes for the development of more classic
10 diagnostic approaches.

This process is defined as gene transcript frequency analysis. The resulting quantitative analysis of the gene transcripts is defined as comparative gene transcript analysis.

15   3. **SUMMARY OF THE INVENTION**
The invention is a method of analyzing a specimen containing gene transcripts comprising the steps of (a) producing a library of biological sequences; (b) generating a set of transcript sequences, where each of the transcript
20 sequences in said set is indicative of a different one of the biological sequences of the library; (c) processing the transcript sequences in a programmed computer (in which a database of reference transcript sequences indicative of reference sequences is stored), to generate an identified
25 sequence value for each of the transcript sequences, where each said identified sequence value is indicative of sequence annotation and a degree of match between one of the biological sequences of the library and at least one of the reference sequences; and (d) processing each said
30 identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

The invention also includes a method of comparing two specimens containing gene transcripts. The first specimen
35 is processed as described above. The second specimen is used to produce a second library of biological sequences, which is used to generate a second set of transcript sequences, where each of the transcript sequences in the

second set is indicative of one of the biological sequences
of the second library. Then the second set of transcript
sequences is processed in a programmed computer to generate
a second set of identified sequence values, namely the

5 further identified sequence values, each of which is
indicative of a sequence annotation and includes a degree
of match between one of the biological sequences of the
second library and at least one of the reference sequences.
The further identified sequence values are processed to

10 generate further final data values indicative of the number
of times each further identified sequence value is present
in the second library. The final data values from the
first specimen and the further identified sequence values
from the second specimen are processed to generate ratios

15 of transcript sequences, which indicate the differences in
the number of gene transcripts between the two specimens.

In a further embodiment, the method includes
quantifying the relative abundance of mRNA in a biological
specimen by (a) isolating a population of mRNA transcripts

20 from a biological specimen; (b) identifying genes from
which the mRNA was transcribed by a sequence-specific
method; (c) determining the numbers of mRNA transcripts
corresponding to each of the genes; and (d) using the mRNA
transcript numbers to determine the relative abundance of

25 mRNA transcripts within the population of mRNA transcripts.

Also disclosed is a method of producing a gene
transcript image analysis by first obtaining a mixture of
mRNA, from which cDNA copies are made. The cDNA is
inserted into a suitable vector which is used to transfect

30 suitable host strain cells which are plated out and
permitted to grow into clones, each cone representing a
unique mRNA. A representative population of clones
transfected with cDNA is isolated. Each clone in the
population is identified by a sequence-specific method

35 which identifies the gene from which the unique mRNA was
transcribed. The number of times each gene is identified
to a clone is determined to evaluate gene transcript
abundance. The genes and their abundances are listed in
order of abundance to produce a gene transcript image.

In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the
5 differences and similarities.

In a further embodiment, the method includes a system for analyzing a library of biological sequences including a means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a
10 different one of the biological sequences of the library; and a means for processing the transcript sequences in a computer system in which a database of reference transcript sequences indicative of reference sequences is stored, wherein the computer is programmed with software for
15 generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and the degree of match between a different one of the biological sequences of the library and at least one of the reference
20 sequences, and for processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

In essence, the invention is a method and system for
25 quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes
30 which are differentially expressed between the two specimens. Thus, this gene transcript image and its comparison can be used as a diagnostic. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs: a
35 gene transcript image. Another embodiment of the method produces the gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, two or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease,

or condition which is correlated to the relative abundance
of gene transcripts in a given cell or population of cells.

4. DESCRIPTION OF THE TABLES AND DRAWINGS
4.1. TABLES

5   Table 1 presents a detailed explanation of the letter
codes utilized in Tables 2-5.

Table 2 lists the one hundred most common gene
transcripts. It is a partial list of isolates from the
HUVEC cDNA library prepared and sequenced as described

10 below. The left-hand column refers to the sequence's order
of abundance in this table. The next column labeled
"number" is the clone number of the first HUVEC sequence
identification reference matching the sequence in the
"entry" column number. Isolates that have not been

15 sequenced are not present in Table 2. The next column,
labeled "N", indicates the total number of cDNAs which have
the same degree of match with the sequence of the reference
transcript in the "entry" column.

The column labeled "entry" gives the NIH GENBANK locus

20 name, which corresponds to the library sequence numbers.
The "s" column indicates in a few cases the species of the
reference sequence. The code for column "s" is given in
Table 1. The column labeled "descriptor" provides a plain
English explanation of the identity of the sequence

25 corresponding to the NIH GENBANK locus name in the "entry"
column.

Table 3 is a comparison of the top fifteen most
abundant gene transcripts in normal monocytes and activated
macrophage cells.

30   Table 4 is a detailed summary of library subtraction
analysis summary comparing the THP-1 and human macrophage
cDNA sequences. In Table 4, the same code as in Table 2 is
used. Additional columns are for "bgfreq" (abundance
number in the subtractant library), "rfend" (abundance

35 number in the target library) and "ratio" (the target
abundance number divided by the subtractant abundance
number). As is clear from perusal of the table, when the
abundance number in the subtractant library is "0", the

target abundance number is divided by 0.05. This is a way
of obtaining a result (not possible dividing by 0) and
distinguishing the result from ratios of subtractant
numbers of 1.

5   Table 5 is the computer program, written in source
code, for generating gene transcript subtraction profiles.
          Table 6 is a partial listing of database entries used
in the electronic northern blot analysis as provided by the
present invention.

10

4.2. BRIEF DESCRIPTION OF THE DRAWINGS

          Figure 1 is a chart summarizing data collected and
stored regarding the library construction portion of
sequence preparation and analysis.

15   Figure 2 is a diagram representing the sequence of
operations performed by "abundance sort" software in a
class of preferred embodiments of the inventive method.
          Figure 3 is a block diagram of a preferred embodiment
of the system of the invention.

20   Figure 4 is a more detailed block diagram of the
bioinformatics process from new sequence (that has already
been sequenced but not identified) to printout of the
transcript imaging analysis and the provision of database
subscriptions.

25   ## 5. DETAILED DESCRIPTION OF THE INVENTION

          The present invention provides a method to compare the
relative abundance of gene transcripts in different
biological specimens by the use of high-throughput
sequence-specific analysis of individual RNAs or their

30 corresponding cDNAs (or alternatively, of data representing
other biological sequences). This process is denoted
herein as gene transcript imaging. The quantitative
analysis of the relative abundance for a set of gene
transcripts is denoted herein as "gene transcript image

35 analysis" or "gene transcript frequency analysis". The
present invention allows one to obtain a profile for gene
transcription in any given population of cells or tissue
from any type of organism. The invention can be applied to

obtain a profile of a specimen consisting of a single cell (or clones of a single cell), or of many cells, or of tissue more complex than a single cell and containing multiple cell types, such as liver.

5   The invention has significant advantages in the fields of diagnostics, toxicology and pharmacology, to name a few. A highly sophisticated diagnostic test can be performed on the ill patient in whom a diagnosis has not been made. A biological specimen consisting of the patient's fluids or

10   tissues is obtained, and the gene transcripts are isolated and expanded to the extent necessary to determine their identity. Optionally, the gene transcripts can be converted to cDNA. A sampling of the gene transcripts are subjected to sequence-specific analysis and quantified.

15   These gene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's data set most closely correlates.

20   For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues, just as it highlights differences between normal monocytes and activated macrophages in Table 3.

In toxicology, a fundamental question is which tests

25   are most effective in predicting or detecting a toxic effect. Gene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. The gene transcript image is a more

30   powerful method to predict drug toxicity and efficacy. Similar benefits accrue in the use of this tool in pharmacology. The gene transcript image can be used selectively to look at protein categories which are expected to be affected, for example, enzymes which

35   detoxify toxins.

In an alternative embodiment, comparative gene transcript frequency analysis is used to differentiate betw en cancer cells which respond to anti-cancer agents and those which do not respond. Examples of anti-cancer

agents are tamoxifen, vincristine, vinblastine,
podophyllotoxins, etoposide, tenisposide, cisplatin,
biologic response modifiers such as interferon, Il-2, GM-
CSF, enzymes, hormones and the like. This method also
5   provides a means for sorting the gene transcripts by
functional category. In the case of cancer cells,
transcription factors or other essential regulatory
molecules are very important categories to analyze across
different libraries.

10   In yet another embodiment, comparative gene transcript
frequency analysis is used to differentiate between control
liver cells and liver cells isolated from patients treated
with experimental drugs like FIAU to distinguish between
pathology caused by the underlying disease and that caused
15 by the drug.

In yet another embodiment, comparative gene transcript
frequency analysis is used to differentiate between brain
tissue from patients treated and untreated with lithium.

In a further embodiment, comparative gene transcript
20 frequency analysis is used to differentiate between
cyclosporin and FK506-treated cells and normal cells.

In a further embodiment, comparative gene transcript
frequency analysis is used to differentiate between virally
infected (including HIV-infected) human cells and
25 uninfected human cells. Gene transcript frequency analysis
is also used to rapidly survey gene transcripts in HIV-
resistant, HIV-infected, and HIV-sensitive cells.
Comparison of gene transcript abundance will indicate the
success of treatment and/or new avenues to study.

30   In a further embodiment, comparative gene transcript
frequency analysis is used to differentiate between
bronchial lavage fluids from healthy and unhealthy patients
with a variety of ailments.

In a further embodiment, comparative gene transcript
35 frequency analysis is used to differentiate between cell,
plant, microbial and animal mutants and wild-type species.
In addition, the transcript abundance program is adapted to
permit the scientist to evaluate the transcription of one
gene in many different tissues. Such comparisons could

identify deletion mutants which do not produce a gene
product and point mutants which produce a less abundant or
otherwise different message. Such mutations can affect
basic biochemical and pharmacological processes, such as

5 mineral nutrition and metabolism, and can be isolated by
means known to those skilled in the art. Thus, crops with
improved yields, pest resistance and other factors can be
developed.

In a further embodiment, comparative gene transcript

10 frequency analysis is used for an interspecies comparative
analysis which would allow for the selection of better
pharmacologic animal models. In this embodiment, humans
and other animals (such as a mouse), or their cultured
cells are treated with a specific test agent. The relative

15 sequence abundance of each cDNA population is determined.
If the animal test system is a good model, homologous genes
in the animal cDNA population should change expression
similarly to those in human cells. If side effects are
detected with the drug, a detailed transcript abundance

20 analysis will be performed to survey gene transcript
changes. Models will then be evaluated by comparing basic
physiological changes.

In a further embodiment, comparative gene transcript
frequency analysis is used in a clinical setting to give a

25 highly detailed gene transcript profile of a patient's
cells or tissue (for example, a blood sample). In
particular, gene transcript frequency analysis is used to
give a high resolution gene expression profile of a
diseased state or condition.

30 In the preferred embodiment, the method utilizes
high-throughput cDNA sequencing to identify specific
transcripts of interest. The generated cDNA and deduced
amino acid sequences are then extensively compared with
GENBANK and other sequence data banks as described below.

35 The method offers several advantages over current protein
discovery by two-dimensional gel methods which try to
identify individual proteins involved in a particular
biological effect. Here, detailed comparisons of profiles
of activated and inactive cells reveal numerous changes in

14

the expression of individual transcripts. After it is
determined if the sequence is an "exact" match, similar or
a non-match, the sequence is entered into a database.
Next, the numbers of copies of cDNA corresponding to each
5   gene are tabulated. Although this can be done slowly and
arduously, if at all, by human hand from a printout of all
entries, a computer program is a useful and rapid way to
tabulate this information. The numbers of cDNA copies
(optionally divided by the total number of sequences in the
10   data set) provides a picture of the relative abundance of
transcripts for each corresponding gene. The list of
represented genes can then be sorted by abundance in the
cDNA population. A multitude of additional types of
comparisons or dimensions are possible and are exemplified
15   below.

     An alternate method of producing a gene transcript
image includes the steps of obtaining a mixture of test
mRNA and providing a representative array of unique probes
whose sequences are complementary to at least some of the
20   test mRNAs. Next, a fixed amount of the test mRNA is added
to the arrayed probes. The test mRNA is incubated with the
probes for a sufficient time to allow hybrids of the test
mRNA and probes to form. The mRNA-probe hybrids are
detected and the quantity determined. The hybrids are
25   identified by their location in the probe array. The
quantity of each hybrid is summed to give a population
number. Each hybrid quantity is divided by the population
number to provide a set of relative abundance data termed a
gene transcript image analysis.

30   6. **EXAMPLES**
     The examples below are provided to illustrate the
subject invention. These examples are provided by way of
illustration and are not included for the purpose of
limiting the invention.

35   6.1. **TISSUE SOURCES AND CELL LINES**
     For analysis with the computer program claimed herein,
biological sequences can be obtained from virtually any

source. Most popular are tissues obtained from the human
body. Tissues can be obtained from any organ of the body,
any age donor, any abnormality or any immortalized cell
line. Immortal cell lines may be preferred in some

5   instances because of their purity of cell type; other
tissue samples invariably include mixed cell types. A
special technique is available to take a single cell (for
example, a brain cell) and harness the cellular machinery
to grow up sufficient cDNA for sequencing by the techniques

10   and analysis described herein (cf. U.S. Patent Nos.
5,021,335 and 5,168,038, which are incorporated by
reference). The examples given herein utilized the
following immortalized cell lines: monocyte-like U-937
cells, activated macrophage-like THP-1 cells, induced

15   vascular endothelial cells (HUVEC cells) and mast cell-like
HMC-1 cells.

The U-937 cell line is a human histiocytic lymphoma
cell line with monocyte characteristics, established from
malignant cells obtained from the pleural effusion of a

20   patient with diffuse histiocytic lymphoma (Sundstrom, C.
and Nilsson, K. (1976) Int. J. Cancer 17:565). U-937 is
one of only a few human cell lines with the morphology,
cytochemistry, surface receptors and monocyte-like
characteristics of histiocytic cells. These cells can be

25   induced to terminal monocytic differentiation and will
express new cell surface molecules when activated with
supernatants from human mixed lymphocyte cultures. Upon
this type of _in vitro_ activation, the cells undergo
morphological and functional changes, including

30   augmentation of antibody-dependent cellular cytotoxicity
(ADCC) against erythroid and tumor target cells (one of the
principal functions of macrophages). Activation of U-937
cells with phorbol 12-myristate 13-acetate (PMA) _in vitro_
stimulates the production of several compounds, including

35   prostaglandins, leukotrienes and platelet-activating factor
(PAF), which are potent inflammatory mediators. Thus, U-
937 is a cell line that is well suited for the
identification and isolation of gene transcripts associated
with normal monocytes.

The HUVEC cell line is a normal, homogeneous, well characterized, early passage endothelial cell culture from human umbilical vein (Cell Systems Corp., 12815 NE 124th Street, Kirkland, WA 98034). Only gene transcripts from
5   induced, or treated, HUVEC cells were sequenced. One batch of 1 X 10^8 cells was treated for 5 hours with 1 U/ml rIL-1b and 100 ng/ml E.coli lipopolysaccharide (LPS) endotoxin prior to harvesting. A separate batch of 2 X 10^8 cells was treated at confluence with 4 U/ml TNF and 2 U/ml
10   interferon-gamma (IFN-gamma) prior to harvesting.

THP-1 is a human leukemic cell line with distinct monocytic characteristics. This cell line was derived from the blood of a 1-year-old boy with acute monocytic leukemia (Tsuchiya, S. et al. (1980) Int. J. Cancer: 171-76). The
15   following cytological and cytochemical criteria were used to determine the monocytic nature of the cell line: 1) the presence of alpha-naphthyl butyrate esterase activity which could be inhibited by sodium fluoride; 2) the production of lysozyme; 3) the phagocytosis of latex particles and
20   sensitized SRBC (sheep red blood cells); and 4) the ability of mitomycin C-treated THP-1 cells to activate T-lymphocytes following ConA (concanavalin A) treatment. Morphologically, the cytoplasm contained small azurophilic granules and the nucleus was indented and irregularly
25   shaped with deep folds. The cell line had Fc and C3b receptors, probably functioning in phagocytosis. THP-1 cells treated with the tumor promoter 12-o-tetradecanoyl-phorbol-13 acetate (TPA) stop proliferating and differentiate into macrophage-like cells which mimic native
30   monocyte-derived macrophages in several respects. Morphologically, as the cells change shape, the nucleus becomes more irregular and additional phagocytic vacuoles appear in the cytoplasm. The differentiated THP-1 cells also exhibit an increased adherence to tissue culture
35   plastic.

HMC-1 cells (a human mast cell line) were established from the peripheral blood of a Mayo Clinic patient with mast cell leukemia (Leukemia Res. (1988) 12:345-55). The cultured cells looked similar to immature cloned murine

mast cells, contained histamine, and stained positively for
chloroacetate esterase, amino caproate esterase, eosinophil
major basic protein (MBP) and tryptase. The HMC-1 cells
have, however, lost the ability to synthesize normal IgE
5   receptors. HMC-1 cells also possess a 10;.16 translocation,
present in cells initially collected by leukophoresis from
the patient and not an artifact of culturing. Thus, HMC-1
cells are a good model for mast cells.

6.2.   CONSTRUCTION OF cDNA LIBRARIES

10   For inter-library comparisons, the libraries must be
prepared in similar manners. Certain parameters appear to
be particularly important to control. One such parameter
is the method of isolating mRNA. It is important to use
the same conditions to remove DNA and heterogeneous nuclear
15   RNA from comparison libraries. Size fractionation of cDNA
must be carefully controlled. The same vector preferably
should be used for preparing libraries to be compared. At
the very least, the same type of vector (e.g.,
unidirectional vector) should be used to assure a valid
20   comparison. A unidirectional vector may be preferred in
order to more easily analyze the output.
It is preferred to prime only with oligo dT
unidirectional primer in order to obtain one only clone per
mRNA transcript when obtaining cDNAs. However, it is
25   recognized that employing a mixture of oligo dT and random
primers can also be advantageous because such a mixture
results in more sequence diversity when gene discovery also
is a goal. Similar effects can be obtained with DR2
(Clontech) and HXLOX (US Biochemical) and also vectors from
30   Invitrogen and Novagen. These vectors have two
requirements. First, there must be primer sites for
commercially available primers such as T3 or M13 reverse
primers. Second, the vector must accept inserts up to 10
kB.
35   It also is important that the clones be randomly
sampled, and that a significant population of clones is
used. Data have been generated with 5,000 clones; however,
if very rare genes are to be obtained and/or their relative

abundance determined, as many as 100,000 clones from a single library may need to be sampled. Size fractionation of cDNA also must be carefully controlled. Alternately, plaques can be selected, rather than clones.

5   Besides the Uni-ZAP™ vector system by Stratagene disclosed below, it is now believed that other similarly unidirectional vectors also can be used. For example, it is believed that such vectors include but are not limited to DR2 (Clontech), and HXLOX (U.S. Biochemical).

10   Preferably, the details of library construction (as shown in Figure 1) are collected and stored in a database for later retrieval relative to the sequences being compared. Fig. 1 shows important information regarding the library collaborator or cell or cDNA supplier,

15 pretreatment, biological source, culture, mRNA preparation and cDNA construction. Similarly detailed information about the other steps is beneficial in analyzing sequences and libraries in depth.

      RNA must be harvested from cells and tissue samples
20 and cDNA libraries are subsequently constructed. cDNA libraries can be constructed according to techniques known in the art. (See, for example, Maniatis, T. et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, New York). cDNA libraries may also be purchased. The U-937

25 cDNA library (catalog No. 937207) was obtained from Stratagene, Inc., 11099 M. Torrey Pines Rd., La Jolla, CA 92037.

      The THP-1 cDNA library was custom constructed by Stratagene from THP-1 cells cultured 48 hours with 100 nm

30 TPA and 4 hours with 1 μg/ml LPS. The human mast cell HMC-1 cDNA library was also custom constructed by Stratagene from cultured HMC-1 cells. The HUVEC cDNA library was custom constructed by Stratagene from two batches of induced HUVEC cells which were separately processed.

35   Essentially, all the libraries were prepared in the same manner. First, poly(A+)RNA (mRNA) was purified. For the U-937 and HMC-1 RNA, cDNA synthesis was only primed with oligo dT. For the THP-1 and HUVEC RNA, cDNA synthesis was primed separately with both oligo dT and random

hexamers, and the two cDNA libraries were treated
separately. Synthetic adaptor oligonucleotides were
ligated onto cDNA ends enabling its insertion into the Uni-
Zap™ vector system (Stratagene), allowing high efficiency
5   unidirectional (sense orientation) lambda library
construction and the convenience of a plasmid system with
blue-white color selection to detect clones with cDNA
insertions. Finally, the two libraries were combined into
a single library by mixing equal numbers of bacteriophage.
10   The libraries can be screened with either DNA probes
or antibody probes and the pBluescript® phagemid
(Stratagene) can be rapidly excised _in vivo_. The phagemid
allows the use of a plasmid system for easy insert
characterization, sequencing, site-directed mutagenesis,
15   the creation of unidirectional deletions and expression of
fusion proteins. The custom-constructed library phage
particles were infected into _E. coli_ host strain XL1-Blue®
(Stratagene), which has a high transformation efficiency,
increasing the probability of obtaining rare, under-
20   represented clones in the cDNA library.

6.3.   ISOLATION OF cDNA CLONES

The phagemid forms of individual cDNA clones were
obtained by the _in vivo_ excision process, in which the host
bacterial strain was coinfected with both the lambda
25   library phage and an f1 helper phage. Proteins derived
from both the library-containing phage and the helper phage
nicked the lambda DNA, initiated new DNA synthesis from
defined sequences on the lambda target DNA and created a
smaller, single stranded circular phagemid DNA molecule
30   that included all DNA sequences of the pBluescript® plasmid
and the cDNA insert. The phagemid DNA was secreted from
the cells and purified, then used to re-infect fresh host
cells, where the double stranded phagemid DNA was produced.
Because the phagemid carries the gene for beta-lactamase,
35   the newly-transformed bacteria are selected on medium
containing ampicillin.
     Phagemid DNA was purified using the Magic Minipreps™
DNA Purification System (Promega catalogue #A7100. Promega

Corp., 2800 Woods Hollow Rd., Madison, WI 53711). This small-scale process provides a simple and reliable method for lysing the bacterial cells and rapidly isolating purified phagemid DNA using a proprietary DNA-binding

5   resin. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

Phagemid DNA was also purified using the QIAwell-8 Plasmid Purification System from QIAGEN® DNA Purification

10   System (QIAGEN Inc., 9259 Eton Ave., Chattsworth, CA 91311). This product line provides a convenient, rapid and reliable high-throughput method for lysing the bacterial cells and isolating highly purified phagemid DNA using QIAGEN anion-exchange resin particles with EMPORE™ membrane

15   technology from 3M in a multiwell format. The DNA was eluted from the purification resin already prepared for DNA sequencing and other analytical manipulations.

An alternate method of purifying phagemid has recently become available. It utilizes the Miniprep Kit (Catalog

20   No. 77468, available from Advanced Genetic Technologies Corp., 19212 Orbit Drive, Gaithersburg, Maryland). This kit is in the 96-well format and provides enough reagents for 960 purifications. Each kit is provided with a recommended protocol, which has been employed except for

25   the following changes. First, the 96 wells are each filled with only 1 ml of sterile terrific broth with carbenicillin at 25 mg/L and glycerol at 0.4%. After the wells are inoculated, the bacteria are cultured for 24 hours and lysed with 60 μl of lysis buffer. A centrifugation step

30   (2900 rpm for 5 minutes) is performed before the contents of the block are added to the primary filter plate. The optional step of adding isopropanol to TRIS buffer is not routinely performed. After the last step in the protocol, samples are transferred to a Beckman 96-well block for

35   storage.

Another new DNA purification system is the WIZARD™ product line which is available from Promega (catalog No. A7071) and may be adaptable to the 96-well format.

21

6.4. SEQUENCING OF cDNA CLONES

The cDNA inserts from random isolates of the U-937 and THP-1 libraries were sequenced in part. Methods for DNA sequencing are well known in the art. Conventional

5   enzymatic methods employ DNA polymerase Klenow fragment, Sequenase™ or Taq polymerase to extend DNA chains from an oligonucleotide primer annealed to the DNA template of interest. Methods have been developed for the use of both single- and double-stranded templates. The chain

10   termination reaction products are usually electrophoresed on urea-acrylamide gels and are detected either by autoradiography (for radionuclide-labeled precursors) or by fluorescence (for fluorescent-labeled precursors). Recent improvements in mechanized reaction preparation, sequencing

15   and analysis using the fluorescent detection method have permitted expansion in the number of sequences that can be determined per day (such as the Applied Biosystems 373 and 377 DNA sequencer, Catalyst 800). Currently with the system as described, read lengths range from 250 to 400

20   bases and are clone dependent. Read length also varies with the length of time the gel is run. In general, the shorter runs tend to truncate the sequence. A minimum of only about 25 to 50 bases is necessary to establish the identification and degree of homology of the sequence.

25   Gene transcript imaging can be used with any sequence-specific method, including, but not limited to hybridization, mass spectroscopy, capillary electrophoresis and 505 gel electrophoresis.

6.5. HOMOLOGY SEARCHING OF cDNA CLONE AND
30   DEDUCED PROTEIN (and Subsequent Steps)

Using the nucleotide sequences derived from the cDNA clones as query sequences (sequences of a Sequence Listing), databases containing previously identified sequences are searched for areas of homology (similarity).

35   Examples of such databases include Genbank and EMBL. We next describe examples of two homology search algorithms that can be used, and then describe the subsequent computer-implemented steps to be performed in accordance with preferred embodiments of the invention.

In the following description of the computer-
implemented steps of the invention, the word "library"
denotes a set (or population) of biological specimen
nucleic acid sequences. A "library" can consist of cDNA
5 sequences, RNA sequences, or the like, which characterize a
biological specimen. The biological specimen can consist
of cells of a single human cell type (or can be any of the
other above-mentioned types of specimens). We contemplate
that the sequences in a library have been determined so as
10 to accurately represent or characterize a biological
specimen (for example, they can consist of representative
cDNA sequences from clones of RNA taken from a single human
cell).

In the following description of the computer-
15 implemented steps of the invention, the expression
"database" denotes a set of stored data which represent a
collection of sequences, which in turn represent a
collection of biological reference materials. For example,
a database can consist of data representing many stored
20 cDNA sequences which are in turn representative of human
cells infected with various viruses, cells of humans of
various ages, cells from different mammalian species, and
so on.

In preferred embodiments, the invention employs a
25 computer programmed with software (to be described) for
performing the following steps:

(a) processing data indicative of a library of cDNA
sequences (generated as a result of high-throughput cDNA
sequencing or other method) to determine whether each
30 sequence in the library matches a DNA sequence of a
reference database of DNA sequences (and if so, identifying
the reference database entry which matches the sequence and
indicating the degree of match between the reference
sequence and the library sequence) and assigning an
35 identified sequence value based on the sequence annotation
and degree of match to each of the sequences in the
library;

(b) for some or all entri s of the database,
tabulating the number of matching identified sequence

values in the library (Although this can be done by human
hand from a printout of all entries, we prefer to perform
this step using computer software to be described below.),
thereby generating a set of final data values or "abundance
5 numbers"; and

(c) if the libraries are different sizes, dividing
each abundance number by the total number of sequences in
the library, to obtain a relative abundance number for each
identified sequence value (i.e., a relative abundance of
10 each gene transcript).

The list of identified sequence values (or genes
corresponding thereto) can then be sorted by abundance in
the cDNA population. A multitude of additional types of
comparisons or dimensions are possible.

15 For example (to be described below in greater detail),
steps (a) and (b) can be repeated for two different
libraries (sometimes referred to as a "target" library and
a "subtractant" library). Then, for each identified
sequence value (or gene transcript), a "ratio" value is
20 obtained by dividing the abundance number (for that
identified sequence value) for the target library, by the
abundance number (for that identified sequence value) for
the subtractant library.

In fact, subtraction may be carried out on multiple
25 libraries. It is possible to add the transcripts from
several libraries (for example, three) and then to divide
them by another set of transcripts from multiple libraries
(again, for example, three). Notation for this operation
may be abbreviated as (A+B+C) / (D+E+F), where the capital
30 letters each indicate an entire library. Optionally the
abundance numbers of transcripts in the summed libraries
may be divided by the total sample size before subtraction.

Unlike standard hybridization technology which permits
a single subtraction of two libraries, once one has
35 processed a set or library transcript sequences and stored
them in the computer, any number of subtractions can be
performed on the library. For example, by this method,
ratio values can be obtained by dividing relative abundance

values in a first library by corresponding values in a
second library and vice versa.

In variations on step (a), the library consists of
nucleotide sequences derived from cDNA clones. Examples of
5   databases which can be searched for areas of homology
(similarity) in step (a) include the commercially available
databases known as Genbank (NIH) EMBL (European Molecular
Biology Labs, Germany), and GENESEQ (Intelligenetics,
Mountain View, California).

10   One homology search algorithm which can be used to
implement step (a) is the algorithm described in the paper
by D.J. Lipman and W.R. Pearson, entitled "Rapid and
Sensitive Protein Similarity Searches," Science, 227:1435
(1985). In this algorithm, the homologous regions are
15   searched in a two-step manner. In the first step, the
highest homologous regions are determined by calculating a
matching score using a homology score table. The parameter
"Ktup" is used in this step to establish the minimum window
size to be shifted for comparing two sequences. Ktup also
20   sets the number of bases that must match to extract the
highest homologous region among the sequences. In this
step, no insertions or deletions are applied and the
homology is displayed as an initial (INIT) value.

In the second step, the homologous regions are aligned
25   to obtain the highest matching score by inserting a gap in
order to add a probable deleted portion. The matching
score obtained in the first step is recalculated using the
homology score Table and the insertion score Table to an
optimized (OPT) value in the final output.

30   DNA homologies between two sequences can be examined
graphically using the Harr method of constructing dot
matrix homology plots (Needleman, S.B. and Wunsch, C.O., J.
Mom. Biol 48:443 (1970)). This method produces a
two-dimensional plot which can be useful in determining
35   regions of homology versus regions of repetition.

However, in a class of preferred embodiments, step (a)
is implemented by processing the library data in the
commercially available computer program known as the
INHERIT 670 Sequence Analysis System, available from

25

Applied Biosystems Inc. (Foster City, California),
including the software known as the Factura software (also
available from Applied Biosystems Inc.). The Factura
program preprocesses each library sequence to "edit out"
5   portions thereof which are not likely to be of interest,
such as the vector used to prepare the library. Additional
sequences which can be edited out or masked (ignored by the
search tools) include but are not limited to the polyA tail
and repetitive GAG and CCC sequences. A low-end search
10 program can be written to mask out such "low-information"
sequences, or programs such as BLAST can ignore the low-
information sequences.

In the algorithm implemented by the INHERIT 670
Sequence Analysis System, the Pattern Specification
15 Language (developed by TRW Inc.) is used to determine
regions of homology. "There are three parameters that
determine how INHERIT analysis runs sequence comparisons:
window size, window offset and error tolerance. Window
size specifies the length of the segments into which the
20 query sequence is subdivided. Window offset specifies
where to start the next segment [to be compared], counting
from the beginning of the previous segment. Error
tolerance specifies the total number of insertions,
deletions and/or substitutions that are tolerated over the
25 specified word length. Error tolerance may be set to any
integer between 0 and 6. The default settings are window
tolerance=20, window offset=10 and error tolerance=3."
INHERIT Analysis Users Manual, pp.2-15. Version 1.0,
Applied Biosystems, Inc., October 1991.
30   Using a combination of these three parameters, a
database (such as a DNA database) can be searched for
sequences containing regions of homology and the
appropriate sequences are scored with an initial value.
Subsequently, these homologous regions are examined using
35 dot matrix homology plots to determine regions of homology
versus regions of repetition. Smith-Waterman alignments
can be used to display the results of the homology search.
The INHERIT software can be executed by a Sun computer
system programmed with the UNIX operating system.

Search alternatives to INHERIT include the BLAST
program, GCG (available from the Genetics Computer Group,
WI) and the Dasher program (Temple Smith, Boston
University, Boston, MA). Nucleotide sequences can be
5   searched against Genbank, EMBL or custom databases such as
GENESEQ (available from Intelligenetics, Mountain View, CA)
or other databases for genes. In addition, we have
searched some sequences against our own in-house database.

In preferred embodiments, the transcript sequences are
10   analyzed by the INHERIT software for best conformance with
a reference gene transcript to assign a sequence identifier
and assigned the degree of homology, which together are the
identified sequence value and are input into, and further
processed by, a Macintosh personal computer (available from
15   Apple) programmed with an "abundance sort and subtraction
analysis" computer program (to be described below).

Prior to the abundance sort and subtraction analysis
program (also denoted as the "abundance sort" program),
identified sequences from the cDNA clones are assigned
20   value (according to the parameters given above) by degree
of match according to the following categories: "exact"
matches (regions with a high degree of identity),
homologous human matches (regions of high similarity, but
not "exact" matches), homologous non-human matches (regions
25   of high similarity present in species other than human), or
non matches (no significant regions of homology to
previously identified nucleotide sequences stored in the
form of the database). Alternately, the degree of match
can be a numeric value as described below.
30   With reference again to the step of identifying
matches between reference sequences and database entries,
protein and peptide sequences can be deduced from the
nucleic acid sequences. Using the deduced polypeptide
sequence, the match identification can be performed in a
35   manner analogous to that done with cDNA sequences. A
protein sequence is used as a query sequence and compared
to the previously identified sequences contained in a
database such as the Swiss/Prot, PIR and the NBRF Protein
database to find homologous proteins. These proteins are

27

initially scored for homology using a homology score Table
(Orcutt, B.C. and Dayoff, M.O. Scoring Matrices, PIR
Report MAT - 0285 (February 1985)) resulting in an INIT
score. The homologous regions are aligned to obtain the
5   highest matching scores by inserting a gap which adds a
probable deleted portion. The matching score is
recalculated using the homology score Table and the
insertion score Table resulting in an optimized (OPT)
score. Even in the absence of knowledge of the proper
10 reading frame of an isolated sequence, the above-described
protein homology search may be performed by searching all 3
reading frames.

Peptide and protein sequence homologies can also be
ascertained using the INHERIT 670 Sequence Analysis System
15 in an analogous way to that used in DNA sequence
homologies. Pattern Specification Language and parameter
windows are used to search protein databases for sequences
containing regions of homology which are scored with an
initial value. Subsequent display in a dot-matrix homology
20 plot shows regions of homology versus regions of
repetition. Additional search tools that are available to
use on pattern search databases include PLsearch Blocks
(available from Henikoff & Henikoff, University of
Washington, Seattle), Dasher and GCG. Pattern search
25 databases include, but are not limited to, Protein Blocks
(available from Henikoff & Henikoff, University of
Washington, Seattle), Brookhaven Protein (available from
the Brookhaven National Laboratory, Brookhaven, MA),
PROSITE (available from Amos Bairoch, University of Geneva,
30 Switzerland), ProDom (available from Temple Smith, Boston
University), and PROTEIN MOTIF FINGERPRINT (available from
University of Leeds, United Kingdom).

The ABI Assembler application software, part of the
INHERIT DNA analysis system (available from Applied
35 Biosystems, Inc., Foster City, CA), can be employed to
create and manage sequence assembly projects by assembling
data from selected sequence fragments into a larger
sequence. The Assembler software combines two advanced
computer technologies which maximize the ability to

assemble sequenced DNA fragments into Assemblages, a
special grouping of data where the relationships betw en
sequences are shown by graphic overlap, alignment and
statistical views. The process is based on the

5 Meyers-Kececioglu model of fragment assembly (INHERIT™
Assembler User's Manual, Applied Biosystems, Inc., Foster
City, CA), and uses graph theory as the foundation of a
very rigorous multiple sequence alignment engine for
assembling DNA sequence fragments. Other assembly programs

10 that can be used include MEGALIGN (available from DNASTAR
Inc., Madison, WI), Dasher and STADEN (available from Roger
Staden, Cambridge, England).

Next, with reference to Fig. 2, we describe in more
detail the "abundance sort" program which implements above-

15 mentioned "step (b)" to tabulate the number of sequences of
the library which match each database entry (the "abundance
number" for each database entry).

Fig. 2 is a flow chart of a preferred embodiment of
the abundance sort program. A source code listing of this

20 embodiment of the abundance sort program is set forth in
Table 5. In the Table 5 implementation, the abundance sort
program is written using the FoxBASE programming language
commercially available from Microsoft Corporation.
Although FoxBASE was the program chosen for the first

25 iteration of this technology, it should not be considered
limiting. Many other programming languages, Sybase being a
particularly desirable alternative, can also be used, as
will be obvious to one with ordinary skill in the art. The
subroutine names specified in Fig. 2 correspond to

30 subroutines listed in Table 5.

With reference again to Fig. 2, the "Identified
Sequences" are transcript sequences representing each
sequence of the library and a corresponding identification
of the database entry (if any) which it matches. In other

35 words, the "Identified Sequences" are transcript sequences
representing the output of above-discussed "step (a)."

Fig. 3 is a block diagram of a system for implementing
the invention. The Fig. 3 system includes library
generation unit 2 which generates a library and asserts an

output stream of transcript sequences indicative of the
biological sequences comprising the library. Programmed
processor 4 receives the data stream output from unit 2 and
processes this data in accordance with above-discussed
5   "step (a)" to generate the Identified Sequences. Processor
4 can be a processor programmed with the commercially
available computer program known as the INHERIT 670
Sequence Analysis System and the commercially available
computer program known as the Factura program (both
10 available from Applied Biosystems Inc.) and with the UNIX
operating system.

      Still with reference to Fig. 3, the Identified
Sequences are loaded into processor 6 which is programmed
with the abundance sort program. Processor 6 generates the
15 Final Transcript sequences indicated in both Figs. 2 and 3.
Fig. 4 shows a more detailed block diagram of a planned
relational computer system, including various searching
techniques which can be implemented, along with an
assortment of databases to query against.

20   With reference to Fig. 2, the abundance sort program
first performs an operation known as "Tempnum" on the
Identified Sequences, to discard all of the Identified
Sequences except those which match database entries of
selected types. For example, the Tempnum process can
25 select Identified Sequences which represent matches of the
following types with database entries (see above for
definition): "exact" matches, human "homologous" matches,
"other species" matches representing genes present in
species other than human), "no" matches (no significant
30 regions of homology with database entries representing
previously identified nucleotide sequences), "I" matches
(Incyte for not previously known DNA sequences), or "X"
matches (matches ESTs in reference database). This
eliminates the U, S, M, V, A, R and D sequence (see Table 1
35 for definitions).

      The identified sequence values selected during the
"Tempnum" process then undergo a further selection (weeding
out) operation known as "Tempred." This operation can, for

example, discard all identified sequence values
representing matches with selected database entries.

The identified sequence values selected during the
"Tempred" process are then classified according to library,
5   during the "Tempdesig" operation. It is contemplated that
the "Identified Sequences" can represent sequences from a
single library, or from two or more libraries.

Consider first the case that the identified sequence
values represent sequences from a single library. In this
10   case, all the identified sequence values determined during
"Tempred" undergo sorting in the "Templib" operation,
further sorting in the "Libsort" operation, and finally
additional sorting in the "Temptarsort" operation. For
example, these three sorting operations can sort the
15   identified sequences in order of decreasing "abundance
number" (to generate a list of decreasing abundance
numbers, each abundance number corresponding to a unique
identified sequence entry, or several lists of decreasing
abundance numbers, with the abundance numbers in each list
20   corresponding to database entries of a selected type) with
redundancies eliminated from each sorted list. In this
case, the operation identified as "Cruncher" can be
bypassed, so that the "Final Data" values are the organized
transcript sequences produced during the "Temptarsort"
25   operation.

We next consider the case that the transcript
sequences produced during the "Tempred" operation represent
sequences from two libraries (which we will denote the
"target" library and the "subtractant" library). For
30   example, the target library may consist of cDNA sequences
from clones of a diseased cell, while the subtractant
library may consist of cDNA sequences from clones of the
diseased cell after treatment by exposure to a drug. For
another example, the target library may consist of cDNA
35   sequences from clones of a cell type from a young human,
while the subtractant library may consist of cDNA sequences
from clones of the same cell type from the same human at
different ages.

In this case, the "Tempdesig" operation routes all transcript sequences representing the target library for processing in accordance with "Templib" (and then "Libsort" and "Temptarsort"), and routes all transcript sequences

5 representing the subtractant library for processing in accordance with "Tempsub" (and then "Subsort" and "Tempsubsort"). For example, the consecutive "Templib," "Libsort," and "Temptarsort" sorting operations sort identified sequences from the target library in order of

10 decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected

15 type) with redundancies eliminated from each sorted list. The consecutive "Tempsub," "Subsort," and "Tempsubsort" sorting operations sort identified sequences from the subtractant library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each

20 abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list.

25   The transcript sequences output from the "Temptarsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of target library sequences), and position along another

30 (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type). Similarly, the transcript sequences output from the "Tempsubsort" operation typically represent sorted lists from which a histogram could be generated in which position along one

35 (e.g., horizontal) axis indicates abundance number (of subtractant library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type).

The transcript sequences (sorted lists) output from
the Tempsubsort and Temptarsort sorting operations are
combined during the operation identified as "Cruncher."
The "Cruncher" process identifies pairs of corresponding
5   target and subtractant abundance numbers (both representing
the same identified sequence value), and divides one by the
other to generate a "ratio" value for each pair of
corresponding abundance numbers, and then sorts the ratio
values in order of decreasing ratio value. The data output
10 from the "Cruncher" operation (the Final Transcript
sequence in Fig. 2) is typically a sorted list from which a
histogram could be generated in which position along one
axis indicates the size of a ratio of abundance numbers
(for corresponding identified sequence values from target
15 and subtractant libraries) and position along another axis
indicates identified sequence value (e.g., gene type).

Preferably, prior to obtaining a ratio between the two
library abundance values, the Cruncher operation also
divides each ratio value by the total number of sequences
20 in one or both of the target and subtractant libraries.
The resulting lists of "relative" ratio values generated by
the Cruncher operation are useful for many medical,
scientific, and industrial applications. Also preferably,
the output of the Cruncher operation is a set of lists,
25 each list representing a sequence of decreasing ratio
values for a different selected subset (e.g. protein
family) of database entries.

In one example, the abundance sort program of the
invention tabulates for a library the numbers of mRNA
30 transcripts corresponding to each gene identified in a
database. These numbers are divided by the total number of
clones sampled. The results of the division reflect the
relative abundance of the mRNA transcripts in the cell type
or tissue from which they were obtained. Obtaining this
35 final data set is referred to herein as "gene transcript
image analysis." The resulting subtracted data show
exactly what proteins and genes are upregulated and
downregulated in highly detailed complexity.

6.6. HUVEC cDNA LIBRARY

Table 2 is an abundance table listing the various gene
transcripts in an induced HUVEC library. The transcripts
are listed in order of decreasing abundance. This

5   computerized sorting simplifies analysis of the tissue and
speeds identification of significant new proteins which are
specific to this cell type. This type of endothelial cell
lines tissues of the cardiovascular system, and the more
that is known about its composition, particularly in

10 response to activation, the more choices of protein targets
become available to affect in treating disorders of this
tissue, such as the highly prevalent atherosclerosis.

6.7.   MONOCYTE-CELL AND MAST-CELL cDNA LIBRARIES

Tables 3 and 4 show truncated comparisons of two

15 libraries. In Tables 3 and 4 the "normal monocytes" are
the HMC-1 cells, and the "activated macrophages" are the
THP-1 cells pretreated with PMA and activated with LPS.
Table 3 lists in descending order of abundance the most
abundant gene transcripts for both cell types. With only

20 15 gene transcripts from each cell type, this table permits
quick, qualitative comparison of the most common
transcripts. This abundance sort, with its convenient
side-by-side display, provides an immediately useful
research tool. In this example, this research tool

25 discloses that 1) only one of the top 15 activated
macrophage transcripts is found in the top 15 normal
monocyte gene transcripts (poly A binding protein); and 2)
a new gene transcript (previously unreported in other
databases) is relatively highly represented in activated

30 macrophages but is not similarly prominent in normal
macrophages. Such a research tool provides researchers
with a short-cut to new proteins, such as receptors, cell-
surface and intracellular signalling molecules, which can
serve as drug targets in commercial drug screening

35 programs. Such a tool could save considerable time over
that consumed by a hit and miss discovery program aimed at
identifying important proteins in and around cells, because
those proteins carrying out everyday cellular functions and

34

represented as steady state mRNA are quickly eliminated
from further characterization.

This illustrates how the gene transcript profiles
change with altered cellular function. Those skilled in
5   the art know that the biochemical composition of cells also
changes with other functional changes such as cancer,
including cancer's various stages, and exposure to
toxicity. A gene transcript subtraction profile such as in
Table 3 is useful as a first screening tool for such gene
10 expression and protein studies.

6.8. SUBTRACTION ANALYSIS OF NORMAL MONOCYTE-CELL AND ACTIVATED MONOCYTE CELL cDNA LIBRARIES

Once the cDNA data are in the computer, the computer
program as disclosed in Table 5 was used to obtain ratios
15 of all the gene transcripts in the two libraries discussed
in Example 6.7, and the gene transcripts were sorted by the
descending values of their ratios. If a gene transcript is
not represented in one library, that gene transcript's
abundance is unknown but appears to be less than 1. As an
20 approximation -- and to obtain a ratio, which would not be
possible if the unrepresented gene were given an abundance
of zero -- genes which are represented in only one of the
two libraries are assigned an abundance of 1/2. Using 1/2
for unrepresented clones increases the relative importance
25 of "turned-on" and "turned-off" genes, whose products would
be drug candidates. The resulting print-out is called a
subtraction table and is an extremely valuable screening
method, as is shown by the following data.

Table 4 is a subtraction table, in which the normal
30 monocyte library was electronically "subtracted" from the
activated macrophage library. This table highlights most
effectively the changes in abundance of the gene
transcripts by activation of macrophages. Even among the
first 20 gene transcripts listed, there are several unknown
35 gene transcripts. Thus, electronic subtraction is a useful
tool with which to assist researchers in identifying much
more quickly the basic biochemical changes between two cell
types. Such a tool can save universities and
pharmaceutical companies which spend billions of dollars on

research valuable time and laboratory resources at the
early discovery stage and can speed up the drug development
cycle, which in turn permits researchers to set up drug
screening programs much earlier. Thus, this research tool
5 provides a way to get new drugs to the public faster and
more economically.

Also, such a subtraction table can be obtained for
patient diagnosis. An individual patient sample (such as
monocytes obtained from a biopsy or blood sample) can be
10 compared with data provided herein to diagnose conditions
associated with macrophage activation.

Table 4 uncovered many new gene transcripts (labeled
Incyte clones). Note that many genes are turned on in the
activated macrophage (i.e., the monocyte had a 0 in the
15 bgfreq column). This screening method is superior to other
screening techniques, such as the western blot, which are
incapable of uncovering such a multitude of discrete new
gene transcripts.

The subtraction-screening technique has also uncovered
20 a high number of cancer gene transcripts (oncogenes rho,
ETS2, rab-2 ras, YPT1-related, and acute myeloid leukemia
mRNA) in the activated macrophage. These transcripts may
be attributed to the use of immortalized cell lines and are
inherently interesting for that reason. This screening
25 technique offers a detailed picture of upregulated
transcripts including oncogenes, which helps explain why
anti-cancer drugs interfere with the patient's immunity
mediated by activated macrophages. Armed with knowledge
gained from this screening method, those skilled in the art
30 can set up more targeted, more effective drug screening
programs to identify drugs which are differentially
effective against 1) both relevant cancers and activated
macrophage conditions with the same gene transcript
profile; 2) cancer alone; and 3) activated macrophage
35 conditions.

Smooth muscle senescent protein (22 kd) was
upregulated in the activated macrophage, which indicates
that it is a candidate to block in controlling
inflammation.

6.9. SUBTRACTION ANALYSIS OF NORMAL LIVER CELLS AND
HEPATITIS INFECTED LIVER CELL cDNA LIBRARIES

In this example, rats are exposed to hepatitis virus
and maintained in the colony until they show definite signs
5   of hepatitis. Of the rats diagnosed with hepatitis, one
half of the rats are treated with a new anti-hepatitis
agent (AHA). Liver samples are obtained from all rats
before exposure to the hepatitis virus and at the end of
AHA treatment or no treatment. In addition, liver samples
10 can be obtained from rats with hepatitis just prior to AHA
treatment.

The liver tissue is treated as described in Examples
6.2 and 6.3 to obtain mRNA and subsequently to sequence
cDNA. The cDNA from each sample are processed and analyzed
15 for abundance according to the computer program in Table 5.
The resulting gene transcript images of the cDNA provide
detailed pictures of the baseline (control) for each animal
and of the infected and/or treated state of the animals.
cDNA data for a group of samples can be combined into a
20 group summary gene transcript profile for all control
samples, all samples from infected rats and all samples
from AHA-treated rats.

Subtractions are performed between appropriate
individual libraries and the grouped libraries. For
25 individual animals, control and post-study samples can be
subtracted. Also, if samples are obtained before and after
AHA treatment, that data from individual animals and
treatment groups can be subtracted. In addition, the data
for all control samples can be pooled and averaged. The
30 control average can be subtracted from averages of both
post-study AHA and post-study non-AHA cDNA samples. If
pre- and post-treatment samples are available, pre- and
post-treatment samples can be compared individually (or
electronically averaged) and subtracted.

35   These subtraction tables are used in two general ways.
First, the differences are analyzed for gene transcripts
which are associated with continuing hepatic deterioration
or healing. The subtraction tables are tools to isolate
the effects of the drug treatment from the underlying basic
40 pathology of hepatitis. Because hepatitis affects many

parameters, additional liver toxicity has been difficult to
detect with only blood tests for the usual enzymes. The
gene transcript profile and subtraction provides a much
more complex biochemical picture which researchers have
5   needed to analyze such difficult problems.

Second, the subtraction tables provide a tool for
identifying clinical markers, individual proteins or other
biochemical determinants which are used to predict and/or
evaluate a clinical endpoint, such as disease, improvement
10 due to the drug, and even additional pathology due to the
drug. The subtraction tables specifically highlight genes
which are turned on or off. Thus, the subtraction tables
provide a first screen for a set of gene transcript
candidates for use as clinical markers. Subsequently,
15 electronic subtractions of additional cell and tissue
libraries reveal which of the potential markers are in fact
found in different cell and tissue libraries. Candidate
gene transcripts found in additional libraries are removed
from the set of potential clinical markers. Then, tests of
20 blood or other relevant samples which are known to lack and
have the relevant condition are compared to validate the
selection of the clinical marker. In this method, the
particular physiologic function of the protein transcript
need not be determined to qualify the gene transcript as a
25 clinical marker.

6.10.   ELECTRONIC NORTHERN BLOT

One limitation of electronic subtraction is that it is
difficult to compare more than a pair of images at once.
Once particular individual gene products are identified as
30 relevant to further study (via electronic subtraction or
other methods), it is useful to study the expression of
single genes in a multitude of different tissues. In the
lab, the technique of "Northern" blot hybridization is used
for this purpose. In this technique, a single cDNA, or a
35 probe corresponding thereto, is labeled and then hybridized
against a blot containing RNA samples prepared from a
multitude of tissues or cell types. Upon autoradiography,

the pattern of expression of that particular gene, one at a
time, can be quantitated in all the included samples.

In contrast, a further embodiment of this invention is
the computerized form of this process, termed here

5   "electronic northern blot." In this variation, a single
gene is queried for expression against a multitude of
prepared and sequenced libraries present within the
database. In this way, the pattern of expression of any
single candidate gene can be examined instantaneously and

10 effortlessly. More candidate genes can thus be scanned,
leading to more frequent and fruitfully relevant
discoveries. The computer program included as Table 5
includes a program for performing this function, and Table
6 is a partial listing of entries of the database used in

15 the electronic northern blot analysis.

6.11. PHASE I CLINICAL TRIALS

Based on the establishment of safety and effectiveness
in the above animal tests, Phase I clinical tests are
undertaken. Normal patients are subjected to the usual

20 preliminary clinical laboratory tests. In addition,
appropriate specimens are taken and subjected to gene
transcript analysis. Additional patient specimens are
taken at predetermined intervals during the test. The
specimens are subjected to gene transcript analysis as

25 described above. In addition, the gene transcript changes
noted in the earlier rat toxicity study are carefully
evaluated as clinical markers in the followed patients.
Changes in the gene transcript analyses are evaluated as
indicators of toxicity by correlation with clinical signs

30 and symptoms and other laboratory results. In addition,
subtraction is performed on individual patient specimens
and on averaged patient specimens. The subtraction
analysis highlights any toxicological changes in the
treated patients. This is a highly refined determinant of

35 toxicity. The subtraction method also annotates clinical
markers. Further subgroups can be analyzed by subtraction
analysis, including, for example, 1) segregation by

occurrence and type of adverse effect; and 2) segregation
by dosage.

6.12. GENE TRANSCRIPT IMAGING ANALYSIS IN CLINICAL STUDIES

A gene transcript imaging analysis (or multiple gene
5 transcript imaging analyses) is a useful tool in other
clinical studies. For example, the differences in gene
transcript imaging analyses before and after treatment can
be assessed for patients on placebo and drug treatment.
This method also effectively screens for clinical markers
10 to follow in clinical use of the drug.

6.13. COMPARATIVE GENE TRANSCRIPT ANALYSIS BETWEEN SPECIES

The subtraction method can be used to screen cDNA
libraries from diverse sources. For example, the same cell
types from different species can be compared by gene
15 transcript analysis to screen for specific differences,
such as in detoxification enzyme systems. Such testing
aids in the selection and validation of an animal model for
the commercial purpose of drug screening or toxicological
testing of drugs intended for human or animal use. When
20 the comparison between animals of different species is
shown in columns for each species, we refer to this as an
interspecies comparison, or zoo blot.

Embodiments of this invention may employ databases
such as those written using the FoxBASE programming
25 language commercially available from Microsoft Corporation.
Other embodiments of the invention employ other databases,
such as a random peptide database, a polymer database, a
synthetic oligomer database, or a oligonucleotide database
of the type described in U.S. Patent 5,270,170, issued
30 December 14, 1993 to Cull, et al., PCT International
Application Publication No. WO 9322684, published November
11, 1993, PCT International Application Publication No. WO
9306121, published April 1, 1993, or PCT International
Application Publication No. WO 9119818, published December
35 26, 1991. These four references (whose text is
incorporated herein by reference) include teaching which

may be applied in implementing such other embodiments of
the present invention.

All references referred to in the preceding text are
hereby expressly incorporated by reference herein.

5   Various modifications and variations of the described
method and system of the invention will be apparent to
those skilled in the art without departing from the scope
and spirit of the invention. Although the invention has
been described in connection with specific preferred

10   embodiments, it should be understood that the invention as
claimed should not be unduly limited to such specific
embodiments.

TABLE 1

Designations (D)

E = Exact
H = Homologous
O = Other species
N = No match
D = Noncoding gene
U = Nonreadable
R = Repetitive DNA
A = Poly-A only
V = Vector only
M = Mitochondrial DNA
S = Skip
I = Match Incyte clone
X = EST match

Distribution (F)

C = Non-specific
P = Cell/tissue specific
U = Unknown

Species (S)

H = Human
A = Ape
P = Pig
D = Dog
V = Bovine
B = Rabbit
R = Rat
M = Mouse
S = Hamster
C = Chicken
F = Amphibian
I = Invertebrate
Z = Protozoan
G = Fungi

Library (L)

U = U937
M = HMC
T = THP-1
H = HUVEC
S = Spleen
L = Lung
Y = T & B cell
A = Adenoid

Localization (Z)

N = Nuclear
C = Cytoplasmic
K = Cytoskeleton
E = Cell surface
Z = Intracellular memb
M = Mitochondrial
S = Secreted
U = Unknown
X = Other

Status (I)

0 = No current interest
1 = Do primary analysis
2 = Primary analysis done
3 = Full length sequence
4 = Secondary analysis
5 = Tissue northern
6 = Obtain full length

Function (R)

T = Translation
L = Protein processing
R = Ribosomal protein
O = Oncogene
G = GTP binding ptn
V = Viral element
Y = Kinase/phosphatase
A = Tumor antigen related
I = Binding proteins
D = NA-binding /transcription
B = Surface molecule/receptor
C = Ca^{++} binding protein
S = Ligands/effectors
H = Stress response protein
E = Enzyme
F = Ferroprotein
P = Protease/inhibitor
Z = Oxidative phosphorylation
Q = Sugar metabolism
M = Amino acid metabolism
N = Nucleic acid metabolism
W = Lipid metabolism
K = Structural
X = Other
U = unknown

4 2

TABLE 2

Clone numbers 15000 through 20000
Libraries: HUVEC
Arranged by ABUNDANCE
Total clones analyzed: 5000

319 genes, for a total of 1713 Clones

	number	N	c	entry	s	descriptor
1	15365	67		HSRPL41		Riboptn L41
2	15004	65		NCY015004		INCYTE 015004
3	15638	63		NCY015638		INCYTE 015638
4	15390	50		NCY015390		INCYTE 015390
5	15193	47		HSFIB1		Fibronectin
6	15220	47		RRRPL9	R	Riboptn L9
7	15280	47		NCY015280		INCYTE 015280
8	15583	33		M62060		EST HHCH09 (IGR)
9	15662	31		HSACTCGR		Actin, gamma
10	15026	29		NCY015026		INCYTE 015026
11	15279	24		HSEF1AR		Elf 1-alpha
12	15027	23		NCY015027		INCYTE 015027
13	15033	20		NCY015033		INCYTE 015033
14	15198	20		NCY015198		INCYTE 015198
15	15809	20		HSCOLL1		Collagenase
16	15221	19		NCY015221		INCYTE 015221
17	15263	19		NCY015263		INCYTE 015263
18	15290	19		NCY015290		INCYTE 015290
19	15350	18		NCY015350		INCYTE 015350
20	15030	17		NCY015030		INCYTE 015030
21	15234	17		NCY015234		INCYTE 015234
22	15459	16		NCY015459		INCYTE 015459
23	15353	15		NCY015353		INCYTE 015353
24	15378	15		S76965		Ptn kinase inhib
25	15255	14		HUMTHYB4		Thymosin beta-4
26	15401	14		HSLIPCR		Lipocortin I
27	15425	14		HSPOLYAB		Poly-A bp
28	18212	14		HUMTHYMA		Thymosin, alpha
29	18216	14		HSMRP1		Motility relat ptn; MRP-1;CD-9
30	15189	13		HS18D		Interferon induc ptn 1-8D
31	15031	12		HUMFKBP		FK506 bp
32	15306	12		HSH2AZ		Histone H2A
33	15621	12		HUMLEC		Lectin, B-galbp, 14kDa
34	15789	11		NCY015789		INCYTE 015789
35	16578	11		HSRPS11		Riboptn S11
36	16632	11		M61984		EST HHCA13 (IGR)
37	18314	11		NCY018314		INCYTE 018314
38	15367	10		NCY015367		INCYTE 015367
39	15415	10		HSIFNIN1		interferon induc mRNA
40	15633	10		HSLDHAR		Lactate dehydrogenase
41	15813	10		CHKNMHCB		C Myosin heavy chain B
42	18210	10		NCY018210		INCYTE 018210
43	18233	10		HSRPII140		RNA polymerase II
44	18996	10		NCY018996		INCYTE 018996
45	15088	9		HUMFERL		Ferritin, light chain
46	15714	9		NCY015714		INCYTE 015714
47	15720	9		NCY015720		INCYTE 015720
48	15863	9		NCY015863		INCYTE 015863
49	16121	9		HSET		Endothelin
50	18252	9		NCY018252		INCYTE 018252
51	15351	8		HUMALBP		Lipid bp, adipocyte
52	15370	8		NCY015370		INCYTE 015370

TABLE 2 Con't

	number	N	c	entry	s	descriptor
53	15670	8		BTCIASHI	V	NADH-ubiq oxidoreductase
54	15795	8		NCY015795		INCYTE 015795
55	16245	8		NCY016245		INCYTE 016245
56	18262	8		NCY018262		INCYTE 018262
57	18321	8		HSRPL17		Riboptn L17
58	15126	7		XLRPL1BRF		Riboptn L1
59	15133	7		HSAC07		Actin, beta
60	15245	7		NCY015245		INCYTE 015245
61	15288	7		NCY015288		INCYTE 015288
62	15294	7		HSGAPDR		G-3-PD
63	15442	7		HUMLAMB		Laminin receptor, 54kDa
64	15485	7		HSNGMRNA		Uracil DNA glycosylase
65	16646	7		NCY016646		INCYTE 016646
66	18003	7		HUMPAIA		Plsmnogen activ gene
67	15032	6		HUMUB		Ubiquitin
68	15267	6		HSRPS8		Riboptn S8
69	15295	6		NCY015295		INCYTE 015295
70	15458	6		RNRPS10R	R	Riboptn S10
71	15832	6		RSGALEM	R	UDP-galactose epimerase
72	15928	6		HUMAPOJ		Apolipoptn J
73	16598	6		HUMTBBM40		Tubulin, beta
74	18218	6		NCY018218		INCYTE 018218
75	18499	6		HSP27		Hydrophobic ptn p27
76	18963	6		NCY018963		INCYTE 018963
77	18997	6		NCY018997		INCYTE 018997
78	15432	5		HSAGALAR		Galactosidase A, alpha
79	15475	5		NCY015475		INCYTE 015475
80	15721	5		NCY015721		INCYTE 015721
81	15865	5		NCY015865		INCYTE 015865
82	16270	5		NCY016270		INCYTE 016270
83	16886	5		NCY016886		INCYTE 016886
84	18500	5		NCY018500		INCYTE 018500
85	18503	5		NCY018503		INCYTE 018503
86	19672	5		RRRPL34	R	Riboptn L34
87	15086	4		XLRPL1AR	F	Riboptn L1a
88	15113	4		HUMIFNWRS		tRNA synthetase, trp
89	15242	4		NCY015242		INCYTE 015242
90	15249	4		NCY015249		INCYTE 015249
91	15377	4		NCY015377		INCYTE 015377
92	15407	4		NCY015407		INCYTE 015407
93	15473	4		NCY015473		INCYTE 015473
94	15588	4		HSRPS12		Riboptn S12
95	15684	4		HSEF1G		Elf 1-gamma
96	15782	4		NCY015782		INCYTE 015782
97	15916	4		HSRPS18		Riboptn S18
98	15930	4		NCY015930		INCYTE 015930
99	16108	4		NCY016108		INCYTE 016108
100	16133	4		NCY016133		INCYTE 016133

44

NORMAL MONONCYTE VS. ACTIVATED MACROPHAGE

Top 15 Most Abundant Genes

NORMAL

1　Elongation factor-1 alpha
2　Ribosomal phosphoprotein
3　Ribosomal protein S8 homolog
4　Beta-Globin
5　Ferritin H chain
6　Ribosomal protein L7
7　Nucleoplasmin
8　Ribosomal protein S20 homolog
9　Transferrin receptor
10　Poly-A binding protein
11　Translationally controlled tumor ptn
12　Ribosomal protein S25
13　Signal recognition particle SRP9
14　Histone H2A.Z
15　Ribosomal protein Ke-3

ACTIVATED

1　Interleukin-1 beta
2　Macrophage inflammatory protein-1
3　Interleukin-8
4　Lymphocyte activation gene
5　Elongation factor-1 alpha
6　Beta actin
7　Rantes T-cell specific protein
8　Poly A binding protein
9　Osteopontin; nephropontin
10　Tumor Necrosis Factor-alpha
11　INCYTE clone 011050
12　Cu/Zn superoxide dismutase
13　Adenylate cyclase (yeast homolog)
14　NGF-related B cell activation molecule
15　Protease Nexin-1, glial-derived

TABLE　3

TABLE 4

Libraries: THP-1
Subtracting: HMC
Sorted by ABUNDANCE
Total clones analyzed: 7375

1057 genes, for a total of 2151 clones

number	entry	s	descriptor	bgfreq	rfend	ratio
10022	HUMIL1		IL 1-beta	0	131	262.00
10036	HSMDNCF		IL-8	0	119	238.00
10089	HSLAG1CDN		Lymphocyte activ gene	0	71	142.00
10060	HUMTCSM		RANTES	0	23	46.000
10003	HUMMIP1A		MIP-1	3	121	40.333
10689	HSOP		Osteopontin	0	20	40.000
11050	NCY011050		INCYTE 011050	0	17	34.000
10937	HSTNFR		TNF-alpha	0	17	34.000
10176	HSSOD		Superoxide dismutase	0	14	28.000
10886	HSCDW40		B-cell activ,NGF-relat	0	10	20.000
10186	HUMAPR		Early resp PMA-induc	0	9	18.000
10967	HUMGDN		PN-1, glial-deriv	0	9	18.000
11353	NCY011353		INCYTE 011353	0	8	16.000
10298	NCY010298		INCYTE 010298	0	7	14.000
10215	HUM4COLA		Collagenase, type IV	0	6	12.000
10276	NCY010276		INCYTE 010276	0	6	12.000
10488	NCY010488		INCYTE 010488	0	6	12.000
11138	NCY011138		INCYTE 011138	0	6	12.000
10037	HUMCAPPRO		Adenylate cyclase	1	10	10.000
10840	HUMADCY		Adenylate cyclase	0	5	10.000
10672	HSCD44E		Cell adhesion glptn	0	5	10.000
12837	HUMCYCLOX		Cyclooxygenase-2	0	5	10.000
10001	NCY010001		INCYTE 010001	0	5	10.000
10005	NCY010005		INCYTE 010005	0	5	10.000
10294	NCY010294		INCYTE 010294	0	5	10.000
10297	NCY010297		INCYTE 010297	0	5	10.000
10403	NCY010403		INCYTE 010403	0	5	10.000
10699	NCY010699		INCYTE 010699	0	5	10.000
10966	NCY010966		INCYTE 010966	0	5	10.000
12092	NCY012092		INCYTE 012092	0	5	10.000
12549	HSRHOB		Oncogene rho	0	5	10.000
10691	HUMARF1BA		ADP-ribosylation fctr	0	4	8.000
12106	HSADSS		Adenylosuccinate synthetase	0	4	8.000
10194	HSCATHL		Cathepsin L	0	4	8.000
10479	CLMCYCA	I	Cyclin A	0	4	8.000
10031	NCY010031		INCYTE 010031	0	4	8.000
10203	NCY010203		INCYTE 010203	0	4	8.000
10288	NCY010288		INCYTE 010288	0	4	8.000
10372	NCY010372		INCYTE 010372	0	4	8.000
10471	NCY010471		INCYTE 010471	0	4	8.000
10484	NCY010484		INCYTE 010484	0	4	8.000
10859	NCY010859		INCYTE 010859	0	4	8.000
10890	NCY010890		INCYTE 010890	0	4	8.000
11511	NCY011511		INCYTE 011511	0	4	8.000
11868	NCY011868		INCYTE 011868	0	4	8.000
12820	NCY012820		INCYTE 012820	0	4	8.000
10133	HSI1RAP		IL-1 antagonist	0	4	8.000
10516	HUMP2A		Phosphatase, regul 2A	0	4	8.000
11063	HUMB94		TNF-induc response	0	4	8.000
11140	HSHB15RNA		HB15 gene; new Ig	0	3	6.000
10788	NCY001713		INCYTE 001713	0	3	6.000
10033	NCY010033		INCYTE 010033	0	3	6.000
10035	NCY010035		INCYTE 010035	0	3	6.000
10084	NCY010084		INCYTE 010084	0	3	6.000
10236	NCY010236		INCYTE 010236	0	3	6.000
10383	NCY010383		INCYTE 010383	0	3	6.000

TABLE 4 Con't

number	entry	s descriptor	bgfreq	rfend	ratio
10450	NCY010450	INCYTE 010450	0	3	6.000
10470	NCY010470	INCYTE 010470	0	3	6.000
10504	NCY010504	INCYTE 010504	0	3	6.000
10507	NCY010507	INCYTE 010507	0	3	6.000
10598	NCY010598	INCYTE 010598	0	3	6.000
10779	NCY010779	INCYTE 010779	0	3	6.000
10909	NCY010909	INCYTE 010909	0	3	6.000
10976	NCY010976	INCYTE 010976	0	3	6.000
10985	NCY010985	INCYTE 010985	0	3	6.000
11052	NCY011052	INCYTE 011052	0	3	6.000
11068	NCY011068	INCYTE 011068	0	3	6.000
11134	NCY011134	INCYTE 011134	0	3	6.000
11136	NCY011136	INCYTE 011136	0	3	6.000
11191	NCY011191	INCYTE 011191	0	3	6.000
11219	NCY011219	INCYTE 011219	0	3	6.000
11386	NCY011386	INCYTE 011386	0	3	6.000
11403	NCY011403	INCYTE 011403	0	3	6.000
11460	NCY011460	INCYTE 011460	0	3	6.000
11618	NCY011618	INCYTE 011618	0	3	6.000
11686	NCY011686	INCYTE 011686	0	3	6.000
12021	NCY012021	INCYTE 012021	0	3	6.000
12025	NCY012025	INCYTE 012025	0	3	6.000
12320	NCY012320	INCYTE 012320	0	3	6.000
12330	NCY012330	INCYTE 012330	0	3	6.000
12853	NCY012853	INCYTE 012853	0	3	6.000
14386	NCY014386	INCYTE 014386	0	3	6.000
14391	NCY014391	INCYTE 014391	0	3	6.000

TABLE 5

```
* Master menu for SUBTRACTION output
SET TALK OFF
SET SAFETY OFF
SET EXACT ON
SET TYPEAHEAD TO 0
CLEAR
SET DEVICE TO SCREEN
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE '          ' TO Target1
STORE '          ' TO Target2
STORE '          ' TO Target3
STORE '          ' TO Object1
STORE '          ' TO Object2
STORE '          ' TO Object3
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO PTF
STORE 1 TO BAIL
DO WHILE .T.
* Program.: Subtraction 2.fmt
* Date....: 10/11/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Subtraction 2
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,24610,-1,8947
@ PIXELS 27,134 SAY "Subtraction Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago";12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,137 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,236 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,35 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,215 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",13 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 90,9 TO 181,109 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 90,288 TO 181,397 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 81,296 SAY "Background:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,135 GET ANAL STYLE 65536 FONT "Chicago",12 PICTURE "@*R Overall;Function" SIZE 4
@ PIXELS 81,26 SAY "Target:" STYLE 65536 FONT "Geneva",270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,20 GET target1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,20 GET target2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,20 GET target3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,299 GET object1 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,299 GET object2 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,299 GET object3 STYLE 0 FONT "Geneva",9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 276,324 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Run;Bail out" SIZE 4112
*
* EOF: Subtraction.2.fmt
READ
   IF Bail=2
   CLEAR
   CLOSE DATABASES
   USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
   SET SAFETY ON
   SCREEN 1 OFF
   RETURN
```

```
   ENDIF
STORE VAL(SYS(2)) TO STARTIME
STORE UPPER(Target1) TO Target1
STORE UPPER(Target2) TO Target2
STORE UPPER(Targ t3) TO Target3
STORE UPPER(Object1) TO Object1
STORE UPPER(Object2) TO Object2
STORE UPPER(Object3) TO Object3
clear
SET TALK ON
GAP = TERMINATE-INITIATE+1
GO INITIATE
COPY NEXT GAP FIELDS NUMBER,library,D,F,Z,R,ENTRY,S,DESCRIPTOR,START,RFEND,I TO TEMPNUM
USE TEMPNUM
COUNT TO TOT
COPY TO TEMPRED FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='I'
USE TEMPRED

IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
COPY TO TEMPDESIG
ELSE
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
   IF Ematch=1
   APPEND FROM TEMPNUM FOR D='E'
   ENDIF
   IF Hmatch=1
   APPEND FROM TEMPNUM FOR D='H'
   ENDIF
   IF Omatch=1
   APPEND FROM TEMPNUM FOR D='O'
   ENDIF
   IF Imatch=1
   APPEND FROM TEMPNUM FOR D='I'.OR.D='X'
*.OR.D='N'
   ENDIF
ENDIF
COUNT TO STARTOT

COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
   APPEND FROM TEMPDESIG FOR library=UPPER(target1)
   IF target2<>'      '
   APPEND FROM TEMPDESIG FOR library=UPPER(target2)
   ENDIF
   IF target3<>'      '
   APPEND FROM TEMPDESIG FOR library=UPPER(target3)
   ENDIF
COUNT TO ANALTOT

USE TEMPDESIG
COPY STRUCTURE TO TEMPSUB
USE TEMPSUB
   APPEND FROM TEMPDESIG FOR library=UPPER(Object1)
   IF target2<>'      '
   APPEND FROM TEMPDESIG FOR library=UPPER(Object2)
   ENDIF
   IF target3<>'      '
   APPEND FROM TEMPDESIG FOR library=UPPER(Object3)
   ENDIF
COUNT TO SUBTRACTOT
SET TALK OFF
******************************************************************************************
* COMPRESSION SUBROUTINE A
? 'COMPRESSING QUERY LIBRARY'
USE TEMPLIB
```

49

```
SORT ON ENTRY,NUMBER TO LIBSORT
USE LIBSORT
COUNT TO IDGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= IDGENE
   PACK
   COUNT TO AUNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
   IF TESTA = TESTB.AND.DESIGA=DESIGB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO.TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPTARSORT
USE TEMPTARSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPTARCO
**********************************************************************************
* COMPRESSION SUBROUTINE B
? 'COMPRESSING TARGET LIBRARY'
USE TEMPSUB
SORT ON ENTRY,NUMBER TO SUBSORT
USE SUBSORT
COUNT TO SUBGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= SUBGENE
   PACK
   COUNT TO BUNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
   IF TESTA = TESTB.AND.DESIGA=DESIGB
```

50

```
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPSUBSORT
USE TEMPSUBSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPSUBCO
**********************************************************************************
*FUSION ROUTINE
? 'SUBTRACTING LIBRARIES'
USE SUBTRACTION
COPY STRUCTURE TO CRUNCHER
SELECT 2
USE TEMPSUBSORT
SELECT 1
USE CRUNCHER
APPEND FROM TEMPTARSORT
COUNT TO BAILOUT
MARK = 0

DO WHILE .T.
SELECT 1
MARK = MARK+1
   IF MARK>BAILOUT
   EXIT
   ENDIF
GO MARK
STORE ENTRY TO SCANNER
SELECT 2
LOCATE FOR ENTRY=SCANNER
IF FOUND()
STORE RFEND TO BIT1
STORE RFEND TO BIT2
ELSE
STORE 1/2 TO BIT1
STORE 0 TO BIT2
ENDIF
SELECT 1
REPLACE BGFREQ WITH BIT2
REPLACE ACTUAL WITH BIT1
LOOP
ENDDO

SELECT 1
REPLACE ALL RATIO WITH RFEND/ACTUAL
? 'DOING FINAL SORT BY RATIO'
SORT ON RATIO/D,BGFREQ/D,DESCRIPTOR TO FINAL
USE FINAL
**********************************************************************************
set talk off
DO CASE
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO 'Adenoid Patent Figures:Subtraction.txt'
```

```
SET ALTERNATE ON
ENDCASE

STORE VAL(SYS(2)) TO FINTIME
IF FINTIME<STARTIME
STORE FINTIME+86400 TO FINTIME
ENDIF
STORE FINTIME - STARTIME TO COMPSEC
STORE COMPSEC/60 TO COMPMIN

*********************
SET MARGIN TO 10
@1,1 SAY 'Library Subtraction Analysis' STYLE 65536 FONT 'Geneva',274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '       '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,5,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
? Target1
IF Target2<>'
?? ','
?? Target2
ENDIF
IF Target3<>'
?? ', '
?? Target3
ENDIF
? 'Subtracting:
? Object1
IF Object2<>'
?? ','
?? Object2
ENDIF
IF Object3<>'
?? ', '
?? Object3
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF ANAL=1
? 'Sorted by ABUNDANCE'
ENDIF
IF ANAL=2
? 'Arranged by FUNCTION'
ENDIF
```

```
? 'Total clones represented: '
?? STR(TOT,5,0)
? 'Total clones analyzed: '
?? STR(STARTOT,5,0)
? 'Total computation time:
?? STR(COMPMIN,5,2)
?? ' minutes'
?
? 'd = designation   f = distribution   z = location   r = function   s = species   i = inte
?
******************************
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',9 COLOR 0,0,0,
DO CASE
CASE ANAL=1
?? STR(AUNIQUE,4,0)
?? ' genes, for a total of '
?? STR(ANALTOT,4,0)
?? ' clones'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I
SET PRINT OFF
CLOSE DATABASES
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'


CASE ANAL=2
* arrange/function
SET PRINT ON
SET HEADING ON
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
?
? '                          BINDING PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Surface molecules and receptors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='B'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Calcium-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='C'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Ligands and effectors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='S'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Other binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFRSQ,RFEND,RATIO,I FOR R='I'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
? '                          ONCOGENES'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'General oncogenes:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='O'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'GTP-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='G'
```

```
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Viral elements:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='V'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Kinases and Phosphatases:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Y'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Tumor-related antigens:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='A'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',268 COLOR 0
? '           PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Transcription and Nucleic Acid-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='D'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Translation:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='T'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Ribosomal proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='R'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Protein processing:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='L'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',268 COLOR 0
?
? '                      ENZYMES'
?

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Ferroproteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='F'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Proteases and inhibitors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='P'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Oxidative phosphorylation:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Z'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Sugar metabolism:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Q'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Amino acid metabolism:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
```

54

```
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='M'
```
```
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Nucleic acid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='N'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Lipid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='W'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other enzymes:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='E'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",268 COLOR 0
?
? '                      MISCELLANEOUS CATEGORIES'
?

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Stress response:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='H'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Structural:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='K'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other clones:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='X'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Clones of unknown function:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='U'
ENDCASE

DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
ERASE TEMPLIB.DBF
ERASE TEMPNUM.DBF
ERASE TEMPDESIG.DBF
SET MARGIN TO 0
CLEAR
LOOP
ENDDO
```

5 5

```
*Northern (single), version 11-25-94
close databas s
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '              ' TO Eobject
STORE '                           ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:  8/ 8/94
* Version.:.FoxBASE+/Mac, revision 1.10
* Notes...:.Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE "SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf"
SET SAFETY OFF
SORT ON library TO 'Compressed libraries.dbf'
* FOR entered>0
SET SAFETY ON
USE 'Compressed libraries.dbf'
DELETE FOR entered=0
PACK
COUNT TO TOT
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
   SW2=1
   LOOP
   ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
SET SAFETY OFF
COPY TO "Hits.dbf" FOR entry=searchval
SET SAFETY ON
```

5 7

```
* MASTER ANALYSIS 3; VERSION 12-9-94
* Master menu for analysis output
CLOSE DATABASES
SET TALK OFF
SET SAFETY OFF
CLEAR
SET DEVICE TO SCREEN
SET DEFAULT TO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:"
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE 0 TO ENTIRE
STORE 0 TO CONDEN
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO XMATCH
STORE 0 TO PRINTON
STORE 0 TO PTF
DO WHILE .T.
* Program.: Master analysis.fmt
* Date....: 12/ 9/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Master analysis
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 39,255 TO 277,430 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 27,98 SAY "Customized Output Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1
@ PIXELS 45,54 GET conden STYLE 65536 FONT "Chicago",12 PICTURE "@*C Condensed format" SIZE
@ PIXELS 54,261 GET anal STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Sort/number;Sort/entry;
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",268 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 63,54 GET PRINTON STYLE 65536 FONT "Chicago",12 PICTURE "@*C Include clone listing"
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,146 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,146 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 234,134 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,125 SAY "->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 189,0 TO 257,120 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 209,8 SAY "Library selection" STYLE 65536 FONT "Geneva",266 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 227,18 GET ENTIRE STYLE 65536 FONT "Chicago",12 PICTURE "@*RV All;Selected" SIZE 16
*
* EOF: Master analysis.fmt
READ
   IF ANAL=9
   CLEAR
   CLOSE DATABASES
   ERASE TEMPMASTER.DBF
   USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
   SET SAFETY ON
   SCREEN 1 OFF
   RETURN
   ENDIF
clear
? INITIATE
? TERMINATE
? CONDEN
? ANAL
```

```
? ematch
? Hmatch
? Omatch
? IMATCH
SET TALK ON
   IF ENTIRE=2
USE "Unique libraries.dbf"
   REPLACE ALL i WITH ' '
   BROWSE FIELDS i,libname,library,total,entered AT 0,0
   ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
*COPY TO TEMPNUM FOR NUMBER>=INITIATE.AND.NUMBER<=TERMINATE
*USE TEMPNUM
COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
   IF ENTIRE=1
   APPEND FROM 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf'
   ENDIF
   IF ENTIRE=2
USE "Unique libraries.dbf"
   COPY TO SELECTED FOR UPPER(i)='Y'
   USE SELECTED
   STORE RECCOUNT() TO STOPIT
   MARK=1
      DO WHILE .T.
      IF MARK>STOPIT
      CLEAR
      EXIT
      ENDIF
      USE SELECTED
      GO MARK
      STORE library TO THISONE
      ? 'COPYING '
      ?? THISONE
      USE TEMPLIB
      APPEND FROM "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf" FOR library=THISONE
      STORE MARK+1 TO MARK
      LOOP
      ENDDO
   ENDIF
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
COUNT TO STARTOT
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
   IF Ematch=0 .AND.. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
   APPEND FROM TEMPLIB
   ENDIF
   IF Ematch=1
   APPEND FROM TEMPLIB FOR D='E'
   ENDIF
   IF Hmatch=1
   APPEND FROM TEMPLIB FOR D='H'
   ENDIF
   IF Omatch=1
   APPEND FROM TEMPLIB FOR D='O'
   ENDIF
   IF Imatch=1
   APPEND FROM TEMPLIB FOR D='I'.OR.D='X'.OR.D='N'
   ENDIF
   IF Xmatch=1
   APPEND FROM TEMPLIB FOR D='X'
   ENDIF
COUNT TO ANALTOT
set talk off
**********************
DO CASE
```

5 9

```
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Total function sort.txt"
*SET ALTERNATE TO "H and O function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance con.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Distribution sort.txt"
*SET ALTERNATE TO "Shear stress HUVEC 1:Clone list.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Location sort.txt"
SET ALTERNATE ON
ENDCASE
********************
IF PRINTON=1
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
ENDIF
?
?
?
?
? date()
?? '      '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '  '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF Xmatch=1
?? 'EST'
```

6 0

```
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
? 'l = library    d = designation   f = distribution    z = location    r = function    c = cer
?
********************************************************************************************
USE TEMPDESIG
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
DO CASE
CASE ANAL=1
* sort/number
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER
DO "COMPRESSION number.PRG"
ELSE
SORT TO TEMP1 ON NUMBER
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR
*list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=2
* sort/DESCRIPTOR
SET HEADING ON
*SORT TO TEMP1 ON DESCRIPTOR,ENTRY,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
*SORT TO TEMP1 ON ENTRY,DESCRIPTOR,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
SORT TO TEMP1 ON ENTRY,START/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
IF CONDEN=1
DO "COMPRESSION entry.PRG"
ELSE
USE TEMP1
 list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF
```

```
CASE ANAL=3
* sort by abundance
SET HEADING ON
SORT TO TEMP1 ON ENTRY,NUMBER for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
DO "COMPRESSION abundance.PRG"

CASE ANAL=4
* sort/interest
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER FOR I>0
DO "COMPRESSION interest.PRG"
ELSE
SORT ON I/D,ENTRY TO TEMP1 FOR I>1
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=5
* arrange/location
SET HEADING ON
STORE 4 TO AMPLIFIER
? 'Nuclear:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoplasmic:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoskeleton:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cell surface:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Intracellular membrane:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Mitochondrial:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compressi n location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
```

```
? 'Secreted;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
EJECT
DO "Output heading.prg"
USE "Analysis location.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '         FUNCTIONAL CLASS                    TOTAL    UNIQUE  NEW  % TOTAL'
?
LIST OFF FIELDS Z,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=6
* arrange/distribution
SET HEADING ON
STORE 3 TO AMPLIFIER
? 'Cell/tissue specific distribution;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Non-specific distribution;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown distribution;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
```

```
EJECT
DO 'Output heading.prg'
USE 'Analysis distribution.dbf'
DO 'Create bargraph.prg'
SET HEADING OFF
? '         FUNCTIONAL CLASS                      TOTAL    UNIQUE   % TOTAL'
?
LIST OFF FIELDS P.NAME,CLONES,GENES,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE 'SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf'
ENDIF


CASE ANAL=7
* arrange/function
SET HEADING ON
STORE 10 TO AMPLIFIER
? '                              BINDING PROTEINS'
?
? 'Surface molecules and receptors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Calcium-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Ligands and effectors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Other binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
*EJECT
? '                              ONCOGENES'
?
? 'General oncogenes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'GTP-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF
? 'Viral elements:'
```

6 4

```
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Kinases and Phosphatases:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Tumor-related antigens:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
? 'Transcription and Nucleic Acid-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Translation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ribosomal proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Protein processing:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg".
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              ENZYMES'
?
? 'Ferroproteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Proteases and inhibit rs:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
```

6 5

```
DO "Normal subroutine 1"
ENDIF
? 'Oxidative phosphorylation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Sugar metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Amino acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Nucleic acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Lipid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other enzymes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                          MISCELLANEOUS CATEGORIES'
?
? 'Stress response:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Structural:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other clones:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression functi n.prg"
ELSE
```

```
DO 'Normal subroutine 1'
ENDIF
? 'Clones of unknown function:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF

IF CONDEN=1
EJECT
*SET DEVICE TO PRINTER
*SET PRINT ON
DO 'Output heading.prg'
***
USE 'Analysis function.dbf'
DO 'Create bargraph.prg'
SET HEADING OFF
***
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
***
? '                                                          TOTAL   TOTAL    NEW       DIST
? '         FUNCTIONAL CLASS                         CLONES  GENES   GENES   FUNCTIONAL CLASS'
? '
***
*LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH,COMPANY
LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF
CASE ANAL=8
DO "Subgroup summary 3.prg"
ENDCASE
DO 'Test print.prg'
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
*ERASE TEMPLIB.DBF
*ERASE TEMPNUM.DBF
*ERASE TEMPDESIG.DBF
*ERASE SELECTED.DBF
CLEAR
LOOP
ENDDO
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   COUNT TO NEWGENES FOR D='H'.OR.D='O'
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE Z TO LOC
USE 'Analysis location.dbf'
LOCATE FOR Z=LOC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? '.clones'
? '                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
  ·ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
·*SET PRINTER ON
SORT ON DATE TO TEMP2
USE TEMP2
?? STR(UNIQUE,4,0)
?? ' genes, for a total of'
?? STR(TOT,4,0)
?? ·' clones'
?
? '                        V Coincidence'
COUNT TO P4 FOR I=4
IF P4>0
? STR(P4,3,0)
?? ' genes with priority = 4 (Secondary analysis:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=4
?
ENDIF
COUNT TO P3 FOR I=3
IF P3>0
? STR(P3,3,0)
?? ' genes with priority = 3 (Full insert sequence:)'
list off fields number.RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=3
?·
ENDIF
COUNT TO P2 FOR I=2.
IF P2>0
? STR(P2,3,0)
?? ' genes with priority = 2 (Primary analysis complete:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=2
?
ENDIF
COUNT TO P1 FOR I=1
IF P1>0
```

```
? STR(P1,3,0)
?? ' genes with priority = 1 (Primary analysis needed:)'
list off fields number,RFEND,L,D,P,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=1
ENDIF


*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON NUMBER TO TEMP2
USE TEMP2

?? STR(UNIQUE,4,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                      V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

7 1

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   COUNT TO NEWGENES FOR D='H'.OR.D='O'
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE R TO FUNC
USE "Analysis function.dbf"
LOCATE FOR P=FUNC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES.
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
SET HEADING ON
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
***
? '                    V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I
***
*SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,
*list off fields RFEND,S,DESCRIPTOR

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

72

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE F TO DIST
USE "Analysis distribution.dbf"
LOCATE FOR P=DIST
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
USE TEMP1
sort on rfend/d to TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
USE TEMP1
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                        V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

74

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf'
COPY TO TEMP1 FOR
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '  Coincidence V        V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

7 5

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '   Coincidence V      V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR;INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

7 6

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
*Lifescan menu; version 8-7-94
SET TALK OFF
set device to screen
CLEAR
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
STORE LUPDATE() TO Update
GO BOTTOM
STORE RECNO() TO cloneno
STORE 6 TO Chooser
DO WHILE .T.
* Program.: Lifeseq menu.fmt
* Date....: 1/11/95
* Version.: FoxBASE+/Mac, revision 1.10
* Notes...: Format file Lifeseq menu
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",268 COLOR 0,0,
@ PIXELS 18,126 TO 77,365 STYLE 28479 COLOR 32767,-25600,-1,-16223,-16721,-15725
@ PIXELS 110,29 TO 188,217 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 45,161 SAY "LIFESEQ" STYLE 65536 FONT 'Geneva',536 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 36,269 SAY "TM" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 63,143 SAY "Molecular Biology Desktop" STYLE 65536 FONT "Helvetica",18 COLOR 0,0,0,
@ PIXELS 90,252 TO 251,467 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 117,270 GET Chooser STYLE 65536 FONT 'Chicago',12 PICTURE "@*RV Transcript profiles
@ PIXELS 135,128 SAY Update STYLE 0 FONT 'Geneva',12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 171,128 SAY cloneno STYLE 0 FONT 'Geneva',12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 135,44 SAY "Last update:" STYLE 65536 FONT 'Geneva',12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,44 SAY "Total clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,296 SAY "v1.30" STYLE 65536 FONT 'Geneva',782 COLOR 0,0,-1,-1,-1,-1
*
* EOF: Lifeseq menu.fmt
READ
DO CASE
CASE Chooser=1
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Master analysis 3.prg"
CASE Chooser=2
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Subtraction 2.prg"
CASE Chooser=3
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Northern (single).prg"
CASE Chooser=4
USE "Libraries.dbf"
BROWSE
CASE Chooser=5
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:See individual clone.prg"
CASE Chooser=6
DO "SmartGuy:FoxBASE+/Mac:fox files:Libraries:Output programs:Menu.prg"
CASE Chooser=7
CLEAR
SCREEN 1 OFF
RETURN
ENDCASE

LOOP
ENDDO
```

7 8

```
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '          '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '   '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations; '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
```

```
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
?
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones!
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '            ' TO Eobject
STORE '                      ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:   8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva";12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,
*
* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'            '
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'                      '
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

83

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE 'SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf'
SET SAFETY OFF
SORT ON library TO "Compressed libraries.dbf"
* FOR entered>0
SET SAFETY ON
USE "Compressed libraries.dbf"
DELETE FOR entered=0
PACK
COUNT TO TOT'
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  SW2=1
  LOOP
  ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE 'SmartGuy:FoxBASE+/Mac:F x files:cl nes.dbf'
SET SAFETY OFF
COPY TO 'Hits.dbf' FOR entry=searchval
SET SAFETY ON
```

84

```
CLOSE DATABASES
SELECT 1
USE "Compressed libraries.dbf"
STORE RECCOUNT() TO Entries
SELECT 2
USE "Hits.dbf"
Mark=1
DO WHILE .T.
SELECT 1
IF Mark>Entries
EXIT
ENDIF
GO MARK
STORE library TO Jigger
SELECT 2
COUNT TO Zog FOR library=Jigger
SELECT 1
REPLACE hits with Zog
Mark=Mark+1
LOOP
ENDDO

SELECT 1
BROWSE FIELDS LIBRARY,LIBNAME,ENTERED,HITS AT 0,0
CLEAR
? 'Enter Y to print:'
WAIT TO PRINSET
IF UPPER(PRINSET)='Y'
SET PRINT ON
CLEAR
EJECT
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",14 COLOR 0,0,0
? 'DATABASE ENTRIES MATCHING ENTRY '
?? Searchval
? DATE()
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
LIST OFF FIELDS library,libname,entered,hits
?
?
SELECT 2
LIST OFF FIELDS NUMBER,LIBRARY,D,S,F,Z,R,ENTRY,DESCRIPTOR,RFSTART,START,RFEND
SET TALK OFF
SET PRINT OFF
ENDIF
CLOSE DATABASES
SET TALK OFF
CLEAR
DO "Test print.prg"
RETURN
```

TABLE　6

library	libname
ADENINB01	Inflamed adenoid
ADRENOR01	Adrenal gland (r)
ADRENOT01	Adrenal gland (T)
AMLBNOT01	AML blast cells (T)
BMARNOT01	Bone marrow
BMARNOT02	Bone marrow (T)
CARDNOT01	Cardiac muscle (T)
CHAONOT01	Chin. hamster ovary
CORNNOT01	Corneal stroma
FIBRAGT01	Fibroblast, AT 5
FIBRAGT02	Fibroblast, AT 30
FIBRANT01	Fibroblast, AT
FIBRNGT01	Fibroblast, uv 5
FIBRNGT02	Fibroblast, uv 30
FIBRNOT01	Fibroblast
FIBRNOT02	Fibroblast, normal
HMC1NOT01	Mast cell line HMC-1
HUVELPB01	HUVEC IFN,TNF,LPS
HUVENOB01	HUVEC control
HUVESTB01	HUVEC shear stress
HYPONOB01	Hypothalamus
KIDNNOT01	Kidney (T)
LIVRNOT01	Liver (T)
LUNGNOT01	Lung (T)
MUSCNOT01	Skeletal muscle (T)
OVIDNOB01	Oviduct
PANCNOT01	Pancreas, normal
PITUNOR01	Pituitary (r)
PITUNOT01	Pituitary (T)
PLACNOB01	Placenta
SINTNOT02	Small intestine (T)
SPLNFET01	Spleen+liver, fetal
SPLNNOT02	Spleen (T)
STOMNOT01	Stomach
SYNORAB01	Rheum. synovium
TBLYNOT01	T + B lymphoblast
TESTNOT01	Testis (T)
THP1NOB01	THP-1 control
THP1PEB01	THP phorbol
THP1PLB01	THP-1 phorbol LPS
U937NOT01	U937, monocytic leuk

number	library	d	s	f	z	r	entry	descriptor	ristart	ristart	rfend
2304	U937NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	0	773
3240	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	370	773
3269	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	371	773
4693	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	470	773
8989	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	327	773
9139	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	375	773

WHAT IS CLAIMED IS:

1.   A method of analyzing a specimen containing gene transcripts, said method comprising the steps of:

(a) producing a library of biological sequences;

5   (b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library;

(c) processing the transcript sequences in a
10   programmed computer in which a database of reference transcript sequences indicative of reference biological sequences is stored, to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence
15   annotation and a degree of match between one of the transcript sequences and at least one of the reference transcript sequences; and

(d) processing each said identified sequence value to generate final data values indicative of a number of times
20   each identified sequence value is present in the library.

2.   The method of claim 1, wherein step (a) includes the steps of:

obtaining a mixture of mRNA;

making cDNA copies of the mRNA;
25   isolating a representative population of clones transfected with the cDNA and producing therefrom the library of biological sequences.

3.   The method of claim 1, wherein the biological sequences are cDNA sequences.

30   4.   The method of claim 1, wherein the biological sequences are RNA sequences.

5.   The method of claim 1, wherein the biological sequences are protein sequences.

6. The method of claim 1, wherein a first value of said degree of match is indicative of an exact match, and a second value of said degree of match is indicative of a non-exact match.

5   7. A method of comparing two specimens containing gene transcripts, said method comprising:
       (a) analyzing a first specimen according to the method of claim 1;
       (b) producing a second library of biological
10 sequences;
       (c) generating a second set of transcript sequences, where each of the transcript sequences in said second set is indicative of a different one of the biological sequences of the second library;
15·   (d) processing the second set of transcript sequences in said programmed computer to generate a second set of identified sequence values known as further identified sequence values, where each of the further identified sequence values is indicative of a sequence annotation and
20 a degree of match between one of the biological sequences of the second library and at least one of the reference sequences;
       (e) processing each said further identified sequence value to generate further final data values indicative of a
25 number of times each further identified sequence value is present in the second library; and
       (f) processing the final data values from the first specimen and the further identified sequence values from the second specimen to generate ratios of transcript
30 sequences, each of said ratio values indicative of differences in numbers of gene transcripts between the two specimens.

       8. A method of quantifying relative abundance of mRNA in a biological specimen, said method comprising the steps
35 of:
       (a) isolating a population of mRNA transcripts from the biological specimen;

8 8

(b)   identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c)   determining numbers of mRNA transcripts corresponding to each of the genes; and

5   (d)   using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

9.   A diagnostic method which comprises producing a gene transcript image, said method comprising the steps of:

10   (a)   isolating a population of mRNA transcripts from a biological specimen;

(b)   identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c)   determining numbers of mRNA transcripts
15 corresponding to each of the genes; and

(d)   using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts, where data determining the relative abundance values of mRNA transcripts is the gene
20 transcript image of the biological specimen.

10.   The method of claim 9, further comprising:

(e)   providing a set of standard normal and diseased gene transcript images; and

(f)   comparing the gene transcript image of the
25 biological specimen with the gene transcript images of step (e) to identify at least one of the standard gene transcript images which most closely approximate the gene transcript image of the biological specimen.

11.   The method of claim 9, wherein the biological
30 specimen is biopsy tissue, sputum, blood or urine.

12.   A method of producing a gene transcript image, said method comprising the steps of

(a)   obtaining a mixture of mRNA;

(b)   making cDNA copies of the mRNA;

(c)   inserting the cDNA into a suitable vector and
using said vector to transfect suitable host strain cells
which are plated out and permitted to grow into clones,
each clone representing a unique mRNA;

5   (d)   isolating a representative population of
recombinant clones;

(e)   identifying amplified cDNAs from each clone in
the population by a sequence-specific method which
identifies gene from which the unique mRNA was transcribed;

10   (f)   determining a number of times each gene is
represented within the population of clones as an
indication of relative abundance; and

(g)   listing the genes and their relative abundance in
order of abundance, thereby producing the gene transcript

15   image.

13.   The method of claim 12, also including the step
of diagnosing disease by:
repeating steps (a) through (g) on biological
specimens from random sample of normal and diseased humans,

20   encompassing a variety of diseases, to produce reference
sets of normal and diseased gene transcript images;
obtaining a test specimen from a human, and producing
a test gene transcript image by performing steps (a)
through (g) on said test specimen;

25   comparing the test gene transcript image with the
reference sets of gene transcript images; and
identifying at least one of the reference gene
transcript images which most closely approximates the test
gene transcript image.

30   14.   A computer system for analyzing a library of
biological sequences, said system including:
means for receiving a set of transcript sequences,
where each of the transcript sequences is indicative of a
different one of the biological sequences of the library;

35   and
means for processing the transcript sequences in the
computer system in which a database of reference transcript

sequences indicative of reference biological sequences is
stored, wherein the computer is programmed with software
for generating an identified sequence value for each of the.
transcript sequences, where each said identified sequence
5   value is indicative of a sequence annotation and a degree
of match between a different one of the biological
sequences of the library and at least one of the reference
transcript sequences, and for processing each said
identified sequence value to generate final data values
10   indicative of a number of times each identified sequence
value is present in the library.

15.   The system of claim 14, also including:
library generation means for producing the library of
biological sequences and generating said set of transcript
15   sequences from said library.

16.   The system of claim 15, wherein the library
generation means includes:
means for obtaining a mixture of mRNA;
means for making cDNA copies of the mRNA;
20   means for inserting the cDNA copies into cells and
permitting the cells to grow into clones;
means for isolating a representative population of the
clones and producing therefrom the library of biological
sequences.

9 1

SYBASE database Structure
Library Preparation

Collaborator
Number
Name
Address
Phone
Fax
Email
Project

cDNA Supplier
Number
Name
Address
Phone
Fax

Cell Supplier
Number
Name
Address
Phone
Fax

cDNA Construction
Number
Prep #
Library code
Supplier
Type
mRNA used
vector
primer
directions
Av size
Date ship
Catalog #
Price
Date rec
Cloning sites
Primary size
Background
Unamp titer
Amp titer
Host strain
Genotypes
Actin
Amplification
Comments

Treatment Link
Treatment name
Culture ID

Biological Source
Number
Tissue
Organ
Gender
Age
Species
Race
Pathology
Disease stage
Tissue weight
Source
Lot
PO
Comments

Culture
Number
Source
Lot
PO
Date
Cell density
Quantity
Protocol
Treatment
Comments

mRNA Prep
Number
Culture
Lot
Date
Lapse
Quantity
Weight
Protocol
RNA yield
mRNA yield
% yield
Modifications
Gel appearance
Comments

Figure 1

Identified Sequences

↓

Tempnum

↓

Tempred

↓

Tempdesig

Templib   Tempsub

↓   ↓

Libsort   Subsort

↓   ↓

Temptorsort   Tempsubsort

→ Cruncher ←

↓

Final Data

Figure 2

Figure 3

Incyte Bioinformatics Process

Figure 4

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US95/01160

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; G06F 15/00

US CL : 435/6; 364/413.02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 364/413.02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS ONLINE, APS, transcript, transcripts, cdan#, mrna#, frequenc?, distribut?, abundanc?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	IntelliGenetics Suite, Release 5.4, Advanced Training Manual, issued January 1993 by IntelliGenetics, Inc. 700 East El Camino Real, Mountain View, California 94040, United States of America, pages (1-6)-(1-19) and (2-9)-(2-14), see entire document.	15 and 16 ---------- 1-14
Y	Science, Volume 252, issued 21 June 1991, M.D. Adams et al, "Complementary DNA sequencing: Expressed sequence tags and human genome project", pages 1651-1656, see entire document.	1-16

X Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 APRIL 1995	0 4 MAY 1995

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	JAMES MARTINELL
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US95/01160

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Nucleic Acids Research, Volume 19, No. 25, issued 1991, E. Hara et al, "Subtractive cDNA cloning using oligo(dT)$_{30}$-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells", pages 7097-7104, see entire document.	1-16
X	Nature Genetics, Volume 2, No. 3, issued November 1992, K. Okubo et al, "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression", pages 173-179, see narrative text portion of entire document.	1, 3
---		----------
Y		2 and 4-16

REPORTS

Axl1p sequence following Ser[200] and occurs within the domain of Axl1p that shows homology with hIDE (14). To delete the complete STE23 sequence and create the sta23Δ::URA3 mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGACCTCAT-TCTTGCTCATTTTGATATTGCTC- TGTAGATTG-TACTGAGAGTGCAC-3'; and 5'-GCTACAAACAGC-GTCGACTTGAATGCCCCGACATCTTCGACTGT-GCGGTATTTCACACCG-3') were used to amplify the URA3 sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, Methods Enzymol. 194, 281 (1991)]. To create the axl1Δ::LEU2 mutation contained on p114, a 5.0-kb Sal I fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb Hpa I-Xho I fragment was replaced with a LEU2 fragment. To construct the sta23Δ::LEU2 allele (a deletion corresponding to 931 amino acids) carried on p153, a LEU2 fragment was used to replace the 2.8-kb Pml I-Ecl136 II fragment of STE23, which occurs within a 6.2-kb Hind III-Bgl II genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.6-kb Bam HI fragment containing MFA1, from pKK16 [K. Kuchler, R. E. Sterne, J. Thorner, EMBO J. 8, 3973 (1989)], was ligated into the Bam HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, Yeast 2, 163 (1986)].

24. J. Chant and I. Herskowitz, Cell 65, 1203 (1991).
25. B. W. Matthews, Acc. Chem. Res. 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thorner, J. Cell Biol. 120, 1203 (1993); R. Kolling and C. P. Hollenberg, EMBO J. 13, 3261 (1994); C. Berkower, D. Loayza, S. Michaelis, Mol. Biol. Cell 5, 1185 (1994).
27. A. Bender and J. R. Pringle, Proc. Natl. Acad. Sci. U.S.A 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, Cell 65, 1213 (1991); S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, ibid., p. 1225; H. O. Park, J. Chant, I. Herskowitz, Nature 365, 269 (1993); J. Chant, Trends Genet. 10, 328 (1994); _____ and J. R. Pringle, J. Cell Biol. 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, ibid., p. 767.
28. G. F. Sprague Jr., Methods Enzymol. 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 sst1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (sta22-1), Y115 (mfa1Δ::LEU2), Y142 (axl1::URA3), Y173 (axl1Δ::LEU2), Y220 (axl1::URA3 sta23Δ::URA3), Y221 (sta23Δ::URA3), Y231 (axl1Δ::LEU2 sta23Δ::LEU2), and Y233 (sta23Δ::LEU2). MATα derivatives of SY2625 included the following strains: Y199 (SY2625 made MATα), Y278 (ste22-1), Y195 (mfa1Δ::LEU2), Y196 (axl1Δ::LEU2), and Y197 (axl1::URA3). The EG123 (MATa leu2 ura3 trp1 can1 his4) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (axl1Δ::LEU2), Y223 (axl1::URA3), Y234 (sta23Δ::LEU2), and Y272 (axl1Δ::LEU2 sta23Δ::LEU2). MATα derivatives of EG123 included the following strains: Y214 (EG123 made MATα) and Y293 (axl1Δ::LEU2). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the axl1 ste23 double mutant strains were created by crossing of the appropriate MATa ste23 and MATα axl1 mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, Yeast 2, 163 (1986)] plasmid containing a 5.5-kb Sal I fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the Bgl II site within AXL1, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DQYPYDVPDYA) (29) between amino acids 854 and 855 of the AXL1 prod-

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb Bam HI-Sst I fragment from pAXL1. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides (axl1-H68A, 5'-GTGCTCACAAAGCGCT-GCCAAACCGGC-3'; axl1-E71A, 5'-AAGAATCAT-GTGCGCACAAAGGTGCGC-3'; and axl1-E71D, 5'-AAGAATCATGTGATCACAAAGGTGCGC-3'). The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb Bam HI-Msc I fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS316 plasmids carrying different AXL1 alleles, p124 (axl1-H68A), p130 (axl1-E71A), and p132 (axl1-E71D). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 Bam HI-Msc I fragment, to generate p161 (axl1-E71A), p162 (axl1-

32

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant Arabidopsis thaliana as a model organism. Arabidopsis possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned Arabidopsis cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an Arabidopsis cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†Present address: Synteni, Palo Alto, CA 94303, USA.
‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total Arabidopsis mRNA (4) by a single round of reverse transcription (5). The Arabidopsis mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (6) and scanned

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the Arabidopsis target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of ~1:50,000. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast TRP4 (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total Arabidopsis mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor HAT4 (8). Fluorescent probes representing mRNA from wild-type and HAT4-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the HAT4 cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of HAT4 mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of HAT4 overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between HAT4-transgenic and wild-type plants (Fig 1, C

Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).
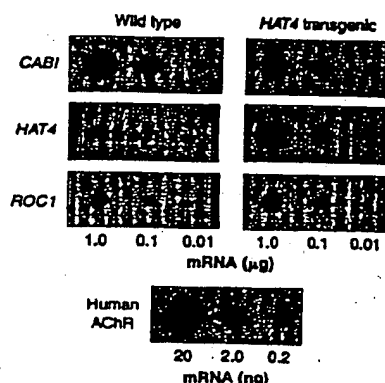

Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR

mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated CABI gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The HAT4-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were

observed for HAT4, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon HAT4 overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and HAT4-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the Arabidopsis genome (2). The availability of 20,274 ESTs from Arabidopsis (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of Arabidopsis genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive in vivo sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	•
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	CABI	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	G44	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	GBF-1	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	GBF-2	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	HAT1	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	HAT2	Homeobox-leucine zipper 2	U09335
e1, 2	HAT4	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	HAT5	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	HAT22	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	KNAT1	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	PPH1	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	ROC1	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	SAR1	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	TRP4	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California).   †No match in the database: novel EST.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, HAT4-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
CABI	1:48	1:83
CABI (tg)	1:120	1:150
HAT4	1:8300	1:6300
HAT4 (tg)	1:150	1:210
ROC1	1:1200	1:1800
ROC1 (tg)	1:260	1:1300

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,645 from the human genome and 21,044 from *Arabidopsis*. Access is available via the World Wide Web (http://www.ncbi.nlm.nih.gov).

2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); I. Hwang et al., *Plant J.* 1, 367 (1991); P. Jarvis et al., *Plant Mol. Biol.* 24, 685 (1994); L. Le Guen et al., *Mol. Gen. Genet.* 245, 390 (1994).

3. D. Shalon, thesis, Stanford University (1995); _____ and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine-coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μl of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited ~0.005 μl per slide on 40 slides at a spacing of 500 μm. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was split according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@cmgm. stanford.edu).

4. F. M. Ausubel et al., Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.

5. Polyadenylated [poly(A)⁺] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50-μl reactions contained 0.1 μg/μl of *Arabidopsis* mRNA, 0.1 ng/μl of human AChR mRNA, 0.05 μg/μl of oligo(dT) (21-mer), 1× first strand buffer, 0.03 U/μl of ribonuclease block, 500 μM deoxyadenosine triphosphate (dATP), 500 μM deoxyguanosine triphosphate, 500 μM dTTP, 40 μM deoxycytidine triphosphate (dCTP), 40 μM fluorescein-12-dCTP (or lissamine-5-dCTP), and 0.03 U/μl of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μl of TE (10 mM tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μl of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μl of 1 M tris-Cl (pH 8.0) and 0.25 μl of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μl of H₂O, and reduced to 3.0 μl in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).

6. Hybridization reactions contained 1.0 μl of fluorescent cDNA synthesis product (5) and 1.0 μl of hybridization buffer [10× saline sodium citrate (SSC) and 0.2% SDS]. The 2.0-μl probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1× SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1× SSC and 0.1% SDS). Arrays were scanned in 0.1× SSC with the use of a fluorescence laser-scanning device (3).

7. Samples of poly(A)⁺ mRNA (4, 5) were spotted onto nylon membranes (Nytran) and crosslinked with ultraviolet light with the use of a Stratalinker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridizations were carried out according to the instructions of the manu-

facturer. Quantitation was performed on a Phosphorimager (Molecular Dynamics).

8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).

9. H. Höfte et al., *Plant J.* 4, 1051 (1993); T. Newman et al., *Plant Physiol.* 106, 1241 (1994).

10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Bellanné-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox et al., *Science* 265, 2031 (1994).

11. E. S. Kawasaki et al., *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).

12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software was developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH (R21HG00450) (P.O.B.) and R37AG00196 (R.W.D.)] and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)–identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6–9).

We previously reported a preclinical model in which ADA gene transfer and expression

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector–mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : G01N 33/543, 33/68	A1	(11) International Publication Number: **WO 95/35505** (43) International Publication Date: 28 December 1995 (28.12.95)

(21) International Application Number: PCT/US95/07659

(22) International Filing Date: 16 June 1995 (16.06.95)

(30) Priority Data:
| 08/261,388 | 17 June 1994 (17.06.94) | US |
| 08/477,809 | 7 June 1995 (07.06.95) | US |

(71) Applicant: THE BOARD OF TRUSTEES OF THE LELAND STANFORD JUNIOR UNIVERSITY [US/US]; Stanford, CA 94305 (US).

(72) Inventors: SHALON, Tidhar, Dari; 364 Fletcher Drive, Atherton, CA 94027 (US). BROWN, Patrick, O.; 76 Peter Coutts Circle, Stanford, CA 94305 (US).

(74) Agent: DEHLINGER, Peter, J.; Dehlinger & Associates, P.O. Box 60850, Palo Alto, CA 94306-1546 (US).

(81) Designated States: AU, CA, JP, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.

(54) Title: METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

(57) Abstract

A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each of a selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.

METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

Field of the Invention

5   This invention relates to a method and apparatus
for fabricating microarrays of biological samples for
large scale screening assays, such as arrays of DNA
samples to be used in DNA hybridization assays for
genetic research and diagnostic applications.

10

References

Abouzied, et al., Journal of AOAC International
77(2):495-500 (1994).

Bohlander, et al., Genomics 13:1322-1324 (1992).

15   Drmanac, et al., Science 260:1649-1652 (1993).

Fodor, et al., Science 251:767-773 (1991).

Khrapko, et al., DNA Sequence 1:375-388 (1991).

Kuriyama, et al., An ISFET BIOSENSOR, APPLIED BIOSENSORS
(Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).

20   Lehrach, et al., HYBRIDIZATION FINGERPRINTING IN GENOME
MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1 (Davies and
Tilgham, Eds.), Cold Spring Harbor Press, pp. 39-81
(1990).

Maniatis, et al., MOLECULAR CLONING, A LABORATORY
25   MANUAL, Cold Spring Harbor Press (1989).

Nelson, et al., Nature Genetics 4:11-18 (1993).

2

Pirrung, et al., U.S. Patent No. 5,143,854 (1992).

Riles, et al., Genetics 134:81-150 (1993).

Schena, M. et al., Proc. Nat. Acad. Sci. USA 89:3894-3898 (1992).

Southern, et al., Genomics 13:1008-1017 (1992).

Background of the Invention

A variety of methods are currently available for making arrays of biological macromolecules, such as arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation. This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22 × 22 cm area (Lehrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable.

3

In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, *et al.* 5   (1992), and also by Fodor, *et al.* (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid 10   sample, *e.g.*, less than 20 bases. A related method has been described by Southern, *et al.* (1992).

Khrapko, *et al.* (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done 15   manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of 20   50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array 25   of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatas, *et al.*, 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled 30   hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for

4

each array which makes the screening and handling of many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of printing horizontal lines of antibodies on a
5   nitrocellulose membrane and separating regions of the membrane with vertical stripes of a hydrophobic material.   Each vertical stripe is then reacted with a different antigen and the reaction between the immobilized antibody and an antigen is detected using a
10   standard ELISA colorimetric technique. Abouzied's technique makes it possible to screen many one-dimensional arrays simultaneously on a single sheet of nitrocellulose.   Abouzied makes the nitrocellulose somewhat hydrophobic using a line drawn with PAP Pen
15   (Research Products International). However Abouzied does not describe a technology that is capable of completely sealing the pores of the nitrocellulose. The pores of the nitrocellulose are still physically open and so the assay reagents can leak through the
20   hydrophobic barrier during extended high temperature incubations or in the presence of detergents which makes the Abouzied technique unacceptable for DNA hybridization assays.

Porous membranes with printed patterns of
25   hydrophilic/hydrophobic regions exist for applications such as ordered arrays of bacteria colonies. QA Life Sciences (San Diego CA) makes such a membrane with a grid pattern printed on it. However, this membrane has the same disadvantage as the Abouzied technique since
30   reagents can still flow between the gridded arrays making them unusable for separate DNA hybridization assays.

Pall Corporation make a 96-well plate with a porous filter heat sealed to the bottom of the plate.
35   These plates are capable of containing different

5

reagents in each well without cross-contamination.
However, each well is intended to hold only one target
element whereas the invention described here makes a
microarray of many biomolecules in each subdivided

5   region of the solid support. Furthermore, the 96 well
plates are at least 1 cm thick and prevent the use of
the device for many colorimetric, fluorescent and
radioactive detection formats which require that the
membrane lie flat against the detection surface. The

10 invention described here requires no further processing
after the assay step since the barriers elements are
shallow and do not interfere with the detection step
thereby greatly increasing convenience.
        Hyseq Corporation has described a method of making

15 an "array of arrays" on a non-porous solid support for
use with their sequencing by hybridization technique.
The method described by Hyseq involves modifying the
chemistry of the solid support material to form a
hydrophobic grid pattern where each subdivided region

20 contains a microarray of biomolecules. Hyseq's flat
hydrophobic pattern does not make use of physical
blocking as an additional means of preventing cross
contamination.

25 **Summary of the Invention**
        The invention includes, in one aspect, a method of
forming a microarray of analyte-assay regions on a
solid support, where each region in the array has a
known amount of a selected, analyte-specific reagent.

30 The method involves first loading a solution of a
selected analyte-specific reagent in a reagent-
dispensing device having an elongate capillary channel
(i) formed by spaced-apart, coextensive elongate
members, (ii) adapted to hold a quantity of the reagent

35 solution and (iii) having a tip region at which aqueous

6

solution in the channel forms a meniscus. The channel
is preferably formed by a pair of spaced-apart tapered
elements.

5   The tip of the dispensing device is tapped against
a solid support at a defined position on the support
surface with an impulse effective to break the meniscus
in the capillary channel deposit a selected volume of
solution on the surface, preferably a selected volume
in the range 0.01 to 100 nl. The two steps are
10   repeated until the desired array is formed.

The method may be practiced in forming a plurality
of such arrays, where the solution-depositing step is
are applied to a selected position on each of a
plurality of solid supports at each repeat cycle.
15   The dispensing device may be loaded with a new
solution, by the steps of (i) dipping the capillary
channel of the device in a wash solution, (ii) removing
wash solution drawn into the capillary channel, and
(iii) dipping the capillary channel into the new
20   reagent solution.

Also included in the invention is an automated
apparatus for forming a microarray of analyte-assay
regions on a plurality of solid supports, where each
region in the array has a known amount of a selected,
25   analyte-specific reagent. The apparatus has a holder
for holding, at known positions, a plurality of planar
supports, and a reagent dispensing device of the type
described above.

The apparatus further includes positioning
30   structure for positioning the dispensing device at a
selected array position with respect to a support in
said holder, and dispensing structure for moving the
dispensing device into tapping engagement against a
support with a selected impulse effective to deposit a

7

selected volume on the support, *e.g.*, a selected volume
in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are
controlled by a control unit in the apparatus. The
5   unit operates to (i) place the dispensing device at a
loading station, (ii) move the capillary channel in the
device into a selected reagent at the loading station,
to load the dispensing device with the reagent, and
(iii) dispense the reagent at a defined array position
10   on each of the supports on said holder. The unit may
further operate, at the end of a dispensing cycle, to
wash the dispensing device by (i) placing the
dispensing device at a washing station, (ii) moving the
capillary channel in the device into a wash fluid, to
15   load the dispensing device with the fluid, and (iii)
remove the wash fluid prior to loading the dispensing
device with a fresh selected reagent.

The dispensing device in the apparatus may be one
of a plurality of such devices which are carried on the
20   arm for dispensing different analyte assay reagents at
selected spaced array positions.

In another aspect, the invention includes a
substrate with a surface having a microarray of at
least 10^3 distinct polynucleotide or polypeptide
25   biopolymers in a surface area of less than about 1 cm^2.
Each distinct biopolymer (i) is disposed at a separate,
defined position in said array, (ii) has a length of at
least 50 subunits, and (iii) is present in a defined
amount between about 0.1 femtomoles and 100 nanomoles.
30   In one embodiment, the surface is glass slide
surface coated with a polycationic polymer, such as
polylysine, and the biopolymers are polynucleotides.
In another embodiment, the substrate has a water-
impermeable backing, a water-permeable film formed on

8

the backing, and a grid formed on the film. The grid
is composed of intersecting water-impervious grid
elements extending from said backing to positions
raised above the surface of said film, and partitions
5   the film into a plurality of water-impervious cells. A
biopolymer array is formed within each well.

More generally, there is provided a substrate for
use in detecting binding of labeled polynucleotides to
one or more of a plurality different-sequence,
10 immobilized polynucleotides. The substrate includes,
in one aspect, a glass support, a coating of a
polycationic polymer, such as polylysine, on said
surface of the support, and an array of distinct
polynucleotides electrostatically bound non-covalently
15 to said coating, where each distinct biopolymer is
disposed at a separate, defined position in a surface
array of polynucleotides.

In another aspect, the substrate includes a water-
impermeable backing, a water-permeable film formed on
20 the backing, and a grid formed on the film, where the
grid is composed of intersecting water-impervious grid
elements extending from the backing to positions raised
above the surface of the film, forming a plurality of
cells. A biopolymer array is formed within each cell.

25 Also forming part of the invention is a method of
detecting differential expression of each of a
plurality of genes in a first cell type, with respect
to expression of the same genes in a second cell type.
In practicing the method, there is first produced
30 fluorescent-labeled cDNA's from mRNA's isolated from
the two cells types, where the cDNA'S from the first
and second cells are labeled with first and second
different fluorescent reporters.

A mixture of the labeled cDNA's from the two cell
35 types is added to an array of polynucleotides

9

representing a plurality of known genes derived fr m
the two cell types, under conditions that result in
hybridization of the cDNA's to complementary-sequence
polynucleotides in the array. The array is then
5   examined by fluorescence under fluorescence excitation
conditions in which (i) polynucleotides in the array
that are hybridized predominantly to cDNA's derived
from one of the first and second cell types give a
distinct first or second fluorescence emission color,
10   respectively, and (ii) polynucleotides in the array
that are hybridized to substantially equal numbers of
cDNA's derived from the first and second cell types
give a distinct combined fluorescence emission color,
respectively. The relative expression of known genes
15   in the two cell types can then be determined by the
observed fluorescence emission color of each spot.
        These and other objects and features of the
invention will become more fully apparent when the
following detailed description of the invention is read
20   in conjunction with the accompanying figures.

Brief Description of the Drawings
        Fig. 1 is a side view of a reagent-dispensing
device having a open-capillary dispensing head
25   constructed for use in one embodiment of the invention;
        Figs. 2A-2C illustrate steps in the delivery of a
fixed-volume bead on a hydrophobic surface employing
the dispensing head from Fig. 1, in accordance with one
embodiment of the method of the invention;
30   Fig. 3 shows a portion of a two-dimensional array
of analyte-assay regions constructed according to the
method of the invention;
        Fig. 4 is a planar view showing components of an
automated apparatus for forming arrays in accordance
35   with the invention.

10

Fig. 5 shows a fluorescent image of an actual 20 × 20 array of 400 fluorescently-labeled DNA samples immobilized on a poly-1-lysine coated slide, where the total area covered by the 400 element array is 16

5   square millimeters;

Fig. 6 is a fluorescent image of a 1.8 cm × 1.8 cm microarray containing lambda clones with yeast inserts, the fluorescent signal arising from the hybridization to the array with approximately half the yeast genome

10   labeled with a green fluorophore and the other half with a red fluorophore;

Fig. 7 shows the translation of the hybridization image of Fig. 6 into a karyotype of the yeast genome, where the elements of Fig.-6 microarray contain yeast

15   DNA sequences that have been previously physically mapped in the yeast genome;

Fig. 8 show a fluorescent image of a 0.5 cm × 0.5 cm microarray of 24 cDNA clones, where the microarray was hybridized simultaneously with total cDNA from wild

20   type Arabidopsis plant labeled with a green fluorophore and total cDNA from a transgenic Arabidopsis plant labeled with a red fluorophore, and the arrow points to the cDNA clone representing the gene introduced into the transgenic Arabidopsis plant;

25   Fig. 9 shows a plan view of substrate having an array of cells formed by barrier elements in the form of a grid;

Fig. 10 shows an enlarged plan view of one of the cells in the substrate in Fig. 9, showing an array of

30   polynucleotide regions in the cell;

Fig. 11 is an enlarged sectional view of the substrate in Fig. 9, taken along a section line in that figure; and

Fig. 12 is a scanned image of a 3 cm × 3 cm

35   nitrocellulose solid support containing four identical

11

arrays of M13 clones in each of four quadrants, where
each quadrant was hybridized simultaneously to a
different oligonucleotide using an open face
hybridization method.

5

Detailed Description of the Invention

I.   Definitions

Unless indicated otherwise, the terms defined
below have the following meanings:

10   "Ligand" refers to one member of a ligand/anti-
ligand binding pair. The ligand may be, for example,
one of the nucleic acid strands in a complementary,
hybridized nucleic acid duplex binding pair; an
effector molecule in an effector/receptor binding pair;
15   or an antigen in an antigen/antibody or
antigen/antibody fragment binding pair.

"Antiligand" refers to the opposite member of a
ligand/anti-ligand binding pair. The antiligand may be
the other of the nucleic acid strands in a
20   complementary, hybridized nucleic acid duplex binding
pair; the receptor molecule in an effector/receptor
binding pair; or an antibody or antibody fragment
molecule in antigen/antibody or antigen/antibody
fragment binding pair, respectively.

25   "Analyte" or "analyte molecule" refers to a
molecule, typically a macromolecule, such as a
polynucleotide or polypeptide, whose presence, amount,
and/or identity are to be determined. The analyte is
one member of a ligand/anti-ligand pair.

30   "Analyte-specific assay reagent" refers to a
molecule effective to bind specifically to an analyte
molecule. The reagent is the opposite member of a
ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a
35   linear or two-dimensional array of pr ferably discrete

12

regi ns, each having a finite area, formed on the surface of a solid support.

A "microarray" is an array of regions having a density of discrete regions of at least about 100/cm², and preferably at least about 1000/cm². The regions in a microarray have typical dimensions, e.g., diameters, in the range of between about 10-250 μm, and are separated from other regions in the array by about the same distance.

A support surface is "hydrophobic" if a aqueous-medium droplet applied to the surface does not spread out substantially beyond the area size of the applied droplet. That is, the surface acts to prevent spreading of the droplet applied to the surface by hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface that forms on the bottom of a liquid in a channel as a result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the biopolymers forming a microarray, means an array member which is distinct from other array members on the basis of a different biopolymer sequence, and/or different concentrations of the same or distinct biopolymers, and/or different mixtures of distinct or different-concentration biopolymers. Thus an array of "distinct polynucleotides" means an array containing, as its members, (i) distinct polynucleotides, which may have a defined amount in each member, (ii) different, graded concentrations of given-sequence polynucleotides, and/or (iii) different-composition mixtures of two or more distinct polynucleotides.

"Cell type" means a cell from a given source, e.g., a tissue, or organ, or a cell in a given state of

13

differentiation, or a cell associated with a given
pathology or genetic makeup.

II. Method of Microarray Formation

5   This section describes a method of forming a
microarray of analyte-assay regions on a solid support
or substrate, where each region in the array has a
known amount of a selected, analyte-specific reagent.
           Fig. 1 illustrates, in a partially schematic view,
10   a reagent-dispensing device 10 useful in practicing the
method. The device generally includes a reagent
dispenser 12 having an elongate open capillary channel
14 adapted to hold a quantity of the reagent solution,
such as indicated at 16, as will be described below.
15   The capillary channel is formed by a pair of spaced-
apart, coextensive, elongate members 12a, 12b which are
tapered toward one another and converge at a tip or tip
region 18 at the lower end of the channel. More
generally, the open channel is formed by at least two
20   elongate, spaced-apart members adapted to hold a
quantity of reagent solutions and having a tip region
at which aqueous solution in the channel forms a
meniscus, such as the concave meniscus illustrated at
20 in Fig. 2A. The advantages of the open channel
25   construction of the dispenser are discussed below.
           With continued reference to Fig. 1, the dispenser
device also includes structure for moving the dispenser
rapidly toward and away from a support surface, for
effecting deposition of a known amount of solution in
30   the dispenser on a support, as will be described below
with reference to Figs. 2A-2C. In the embodiment
shown, this structure includes a solenoid 22 which is
activatable to draw a solenoid piston 24 rapidly
downwardly, then release the piston, e.g., under spring
35   bias, to a normal, raised position, as shown. The

14

dispenser is carried on the piston by a connecting
member 26, as shown. The just-described moving
structure is also referred to herein as dispensing
means for moving the dispenser into engagement with a
5   solid support, for dispensing a known volume of fluid
on the support.

        The dispensing device just described is carried on
an arm 28 that may be moved either linearly or in an x-
y plane to position the dispenser at a selected
10   deposition position, as will be described.

        Figs. 2A-2C illustrate the method of depositing a
known amount of reagent solution in the just-described
dispenser on the surface of a solid support, such as
the support indicated at 30. The support is a polymer,
15   glass, or other solid-material support having a surface
indicated at 31.

        In one general embodiment, the surface is a
relatively hydrophilic, *i.e.*, wettable surface, such as
a surface having native, bound or covalently attached
20   charged groups. On such surface described below is a
glass surface having an absorbed layer of a
polycationic polymer, such as poly-l-lysine.

        In another embodiment, the surface has or is
formed to have a relatively hydrophobic character,
25   *i.e.*, one that causes aqueous medium deposited on the
surface to bead. A variety of known hydrophobic
polymers, such as polystyrene, polypropylene, or
polyethylene have desired hydrophobic properties, as do
glass and a variety of lubricant or other hydrophobic
30   films that may be applied to the support surface.

        Initially, the dispenser is loaded with a selected
analyte-specific reagent solution, such as by dipping
the dispenser tip, after washing, into a solution of
the reagent, and allowing filling by capillary flow
35   into the dispenser channel. The dispenser is now moved

15

to a selected position with respect to a support
surface, placing the dispenser tip directly above the
support-surface position at which the reagent is to be
deposited. This movement takes place with the

5   dispenser tip in its raised position, as seen in Fig.
2A, where the tip is typically at least several 1-5 mm
above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is
now activated to cause the dispenser tip to move

10 rapidly toward and away from the substrate surface,
making momentary contact with the surface, in effect,
tapping the tip of the dispenser against the support
surface. The tapping movement of the tip against the
surface acts to break the liquid meniscus in the tip

15 channel, bringing the liquid in the tip into contact
with the support surface. This, in turn, produces a
flowing of the liquid into the capillary space between
the tip and the surface, acting to draw liquid out of
the dispenser channel, as seen in Fig. 2B.

20 Fig. 2C shows flow of fluid from the tip onto the
support surface, which in this case is a hydrophobic
surface. The figure illustrates that liquid continues
to flow from the dispenser onto the support surface
until it forms a liquid bead 32. At a given bead size,

25 i.e., volume, the tendency of liquid to flow onto the
surface will be balanced by the hydrophobic surface
interaction of the bead with the support surface, which
acts to limit the total bead area on the surface, and
by the surface tension of the droplet, which tends

30 toward a given bead curvature. At this point, a given
bead volume will have formed, and continued contact of
the dispenser tip with the bead, as the dispenser tip
is being withdrawn, will have little or no effect on
bead volume.

16

For liquid-dispensing on a more hydrophilic
surface, the liquid will have less of a tendency to
bead, and the dispensed volume will be more sensitive
to the total dwell time of the dispenser tip in the
5   immediate vicinity of the support surface, e.g., the
positions illustrated in Figs. 2B and 2C.

The desired deposition volume, i.e., bead volume,
formed by this method is preferably in the range 2 pl
(picoliters) to 2 nl (nanoliters), although volumes as
10   high as 100 nl or more may be dispensed. It will be
appreciated that the selected dispensed volume will
depend on (i) the "footprint" of the dispenser tip,
i.e., the size of the area spanned by the tip, (ii) the
hydrophobicity of the support surface, and (iii) the
15   time of contact with and rate of withdrawal of the tip
from the support surface. In addition, bead size may
be reduced by increasing the viscosity of the medium,
effectively reducing the flow time of liquid from the
dispenser onto the support surface. The drop size may
20   be further constrained by depositing the drop in a
hydrophilic region surrounded by a hydrophobic grid
pattern on the support surface.

In a typical embodiment, the dispenser tip is
tapped rapidly against the support surface, with a
25   total residence time in contact with the support of
less than about 1 msec, and a rate of upward travel
from the surface of about 10 cm/sec.

Assuming that the bead that forms on contact with
the surface is a hemispherical bead, with a diameter
30   approximately equal to the width of the dispenser tip,
as shown in Fig. 2C, the volume of the bead formed in
relation to dispenser tip width (d) is given in Table 1
below. As seen, the volume of the bead ranges between
2 pl to 2 nl as the width size is increased from about
35   20 to 200 μm.

17

Tabl 1

d	Volume (nl)
20 μm	2×10^{-3}
50 μm	3.1×10^{-2}
100 μm	2.5×10^{-1}
200 μm	2

At a given tip size, bead volume can be reduced in a controlled fashion by increasing surface hydrophobicity, reducing time of contact of the tip with the surface, increasing rate of movement of the tip away from the surface, and/or increasing the viscosity of the medium. Once these parameters are fixed, a selected deposition volume in the desired pl to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location on a support, the tip is typically moved to a corresponding position on a second support, a droplet is deposited at that position, and this process is repeated until a liquid droplet of the reagent has been deposited at a selected position on each of a plurality of supports.

The tip is then washed to remove the reagent liquid, filled with another reagent liquid and this reagent is now deposited at each another array position on each of the supports. In one embodiment, the tip is washed and refilled by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

From the foregoing, it will be appreciated that the tweezers-like, open-capillary dispenser tip

18

provides the advantages that (i) the open channel of
the tip facilitates rapid, efficient washing and drying
before reloading the tip with a new reagent, (ii)
passive capillary action can load the sample directly

5   from a standard microwell plate while retaining
sufficient sample in the open capillary reservoir for
the printing of numerous arrays, (iii) open capillaries
are less prone to clogging than closed capillaries, and
(iv) open capillaries do not require a perfectly faced

10   bottom surface for fluid delivery.
        A portion of a microarray 36 formed on the surface
38 of a solid support 40 in accordance with the method
just described is shown in Fig. 3. The array is formed
of a plurality of analyte-specific reagent regions,

15   such as regions 42, where each region may include a
different analyte-specific reagent. As indicated
above, the diameter of each region is preferably
between about 20-200 μm. The spacing between each
region and its closest (non-diagonal) neighbor,

20   measured from center-to-center (indicated at 44), is
preferably in the range of about 20-400 μm. Thus, for
example, an array having a center-to-center spacing of
about 250 μm contains about 40 regions/cm or 1,600
regions/cm^2. After formation of the array, the support

25   is treated to evaporate the liquid of the droplet
forming each region, to leave a desired array of dried,
relatively flat regions. This drying may be done by
heating or under vacuum.
        In some cases, it is desired to first rehydrate

30   the droplets containing the analyte reagents to allow
for more time for adsorption to the solid support. It
is also possible to spot out the analyte reagents in a
humid environment so that droplets do not dry until the
arraying operation is complete.

19

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an automated apparatus for forming an array of analyte-assay regions on a solid support, where each region in the array has a known amount of a selected, analyte-specific reagent.

The apparatus is shown in planar, and partially schematic view in Fig. 4. A dispenser device 72 in the apparatus has the basic construction described above with respect to Fig. 1, and includes a dispenser 74 having an open-capillary channel terminating at a tip, substantially as shown in Figs. 1 and 2A-2C.

The dispenser is mounted in the device for movement toward and away from a dispensing position at which the tip of the dispenser taps a support surface, to dispense a selected volume of reagent solution, as described above. This movement is effected by a solenoid 76 as described above. Solenoid 76 is under the control of a control unit 77 whose operation will be described below. The solenoid is also referred to herein as dispensing means for moving the device into tapping engagement with a support, when the device is positioned at a defined array position with respect to that support.

The dispenser device is carried on an arm 74 which is threadedly mounted on a worm screw 80 driven (rotated) in a desired direction by a stepper motor 82 also under the control of unit 77. At its left end in the figure screw 80 is carried in a sleeve 84 for rotation about the screw axis. At its other end, the screw is mounted to the drive shaft of the stepper motor, which in turn is carried on a sleeve 86. The dispenser device, worm screw, the two sleeves mounting the worm screw, and the stepper motor used in moving the device in the "x" (horizontal) direction in the

20

figure form what is referred to here collectively as a
displacement assembly 86.

    The displacement assembly is constructed to
produce precise, micro-range movement in the direction

5   of the screw, i.e., along an x axis in the figure. In
one mode, the assembly functions to move the dispenser
in x-axis increments having a selected distance in the
range 5-25 μm. In another mode, the dispenser unit may
be moved in precise x-axis increments of several

10   microns or more,: for positioning the dispenser at
associated positions on adjacent supports, as will be
described below.

    The displacement assembly, in turn, is mounted for
movement in the "y" (vertical) axis of the figure, for

15   positioning the dispenser at a selected y axis
position. The structure mounting the assembly includes
a fixed rod 88 mounted rigidly between a pair of frame
bars 90, 92, and a worm screw 94 mounted for rotation
between a pair of frame bars 96, 98. The worm screw is

20   driven (rotated) by a stepper motor 100 which operates
under the control of unit 77. The motor is mounted on
bar 96, as shown.

    The structure just described, including worm screw
94 and motor 100, is constructed to produce precise,

25   micro-range movement in the direction of the screw,
i.e., along an y axis in the figure. As above, the
structure functions in one mode to move the dispenser
in y-axis increments having a selected distance in the
range 5-250 μm, and in a second mode, to move the

30   dispenser in precise y-axis increments of several
microns (μm) or more, for positioning the dispenser at
associated positions on adjacent supports.

    The displacement assembly and structure for moving
this assembly in the y axis are referred to herein

35   collectively as positioning means for positioning the

dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which the microarrays of regent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of the supports. Solenoid actuati n of the dispenser is

22

then effective to dispense a selected-volume droplet of
that reagent at this location. As noted above, this
operation is effective to dispense a selected volume
preferably between 2 pl and 2 nl of the reagent
5   solution.

The dispenser is now moved to the corresponding
position at an adjacent support and a similar volume of
the solution is dispensed at this position. The
process is repeated until the reagent has been
10   dispensed at this preselected corresponding position on
each of the supports.

Where it is desired to dispense a single reagent
at more than two array positions on a support, the
dispenser may be moved to different array positions at
15   each support, before moving the dispenser to a new
support, or solution can be dispensed at individual
positions on each support, at one selected position,
then the cycle repeated for each new array position.

To dispense the next reagent, the dispenser is
20   positioned over a wash solution (not shown), and the
dispenser tip is dipped in and out of this solution
until the reagent solution has been substantially
washed from the tip. Solution can be removed from the
tip, after each dipping, by vacuum, compressed air
25   spray, sponge, or the like.

The dispenser tip is now dipped in a second
reagent well, and the filled tip is moved to a second
selected array position in the first support. The
process of dispensing reagent at each of the
30   corresponding second-array positions is then carried as
above. This process is repeated until an entire
microarray of reagent solutions on each of the supports
has been formed.

35   IV.   Microarray Substrate

23

This section describes embodiments of a substrate
having a microarray of biological polymers carried on
the substrate surface. Subsection A describes a multi-
cell substrate, each cell of which contains a
5   microarray, and preferably an identical microarray, of
distinct biopolymers, such as distinct polynucleotides,
formed on a porous surface. Subsection B describes a
microarray of distinct polynucleotides bound on a glass
slide coated with a polycationic polymer.
10

       A.   Multi-Cell Substrate
       Fig. 9 illustrates, in plan view, a substrate 110
constructed according to the invention. The substrate
has an 8 × 12 rectangular array 112 of cells, such as
15   cells 114, 116, formed on the substrate surface. With
reference to Fig. 10, each cell, such as cell 114, in
turn supports a microarray 118 of distinct biopolymers,
such as polypeptides or polynucleotides at known,
addressable regions of the microarray. Two such
20   regions forming the microarray are indicated at 120,
and correspond to regions, such as regions 42, forming
the microarray of distinct biopolymers shown in Fig. 3.
       The 96-cell array shown in Fig. 9 has typically
array dimensions between about 12 and 244 mm in width
25   and 8 and 400 mm in length, with the cells in the array
having width and length dimension of 1/12 and 1/8 the
array width and length dimensions, respectively, i.e.,
between about 1 and 20 in width and 1 and 50 mm in
length.
30   The construction of substrate is shown cross-
sectionally in Fig. 11, which is an enlarged sectional
view taken along view line 124 in Fig. 9. The
substrate includes a water-impermeable backing 126,
such as a glass slide or rigid polymer sheet. Formed
35   on the surface of the backing is a water-permeable film

24

128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is preferably between about 10 and 1000 μm. The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to Fig. 11, the film-covered surface in the substrate is partitioned into a desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of the solid support.

In alternative embodiments, the barrier element
can be a wax-based material or a thermoset material
such as epoxy. The barrier material can also be a UV-
curing polymer which is exposed to UV light after being
5   printed onto the solid support. The barrier material
may also be applied to the solid support using printing
techniques such as silk-screen printing. The barrier
material may also be a heat-seal stamping of the porous
solid support which seals its pores and forms a water-
10 impervious barrier element. The barrier material may
also be a shallow grid which is laminated or otherwise
adhered to the solid support.

In addition to plastic-backed nitrocellulose, the
solid support can be virtually any porous membrane with
15 or without a non-porous backing. Such membranes are
readily available from numerous vendors and are made
from nylon, PVDF, polysulfone and the like. In an
alternative embodiment, the barrier element may also be
used to adhere the porous membrane to a non-porous
20 backing in addition to functioning as a barrier to
prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support
can be of a non-porous material. The barrier can be
printed either before or after the microarray of
25 biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the
grid lines and the underlying backing are water-
impermeable, having side barriers projecting above the
porous film in the cells. Thus, defined-volume samples
30 can be placed in each well without risk of cross-
contamination with sample material in adjacent cells.
In Fig. 11, defined volumes samples, such as sample
134, are shown in the cells.

As noted above, each well contains a microarray of
35 distinct biopolymers. In one general embodiment, the

26

microarrays in the well are identical arrays of
distinct biopolymers, e.g., different sequence
polynucleotides. Such arrays can be formed in
accordance with the methods described in Section II, by
5   depositing a first selected polynucleotide at the same
selected microarray position in each of the cells, then
depositing a second polynucleotide at a different
microarray position in each well, and so on until a
complete, identical microarray is formed in each cell.
10   In a preferred embodiment, each microarray
contains about 10^3 distinct polynucleotide or
polypeptide biopolymers per surface area of less than
about 1 cm^2. Also in a preferred embodiment, the
biopolymers in each microarray region are present in a
15   defined amount between about 0.1 femtomoles and 100
nanomoles. The ability to form high-density arrays of
biopolymers, where each region is formed of a well-
defined amount of deposited material, can be achieved
in accordance with the microarray-forming method
20   described in Section II.
         Also in a preferred embodiments, the biopolymers
are polynucleotides having lengths of at least about 50
bp, i.e., substantially longer than oligonucleotides
which can be formed in high-density arrays by schemes
25   involving parallel, step-wise polymer synthesis on the
array surface.
         In the case of a polynucleotide array, in an assay
procedure, a small volume of the labeled DNA probe
mixture in a standard hybridization solution is loaded
30   onto each cell. The solution will spread to cover the
entire microarray and stop at the barrier elements.
The solid support is then incubated in a humid chamber
at the appropriate temperature as required by the
assay.

27

Each assay may be conducted in an "open-face"
format where no further sealing step is required, since
the hybridization solution will be kept properly
hydrated by the water vapor in the humid chamber.   At
5   the conclusion of the incubation step, the entire solid
support containing the numerous microarrays is rinsed
quickly enough to dilute the assay reagents so that no
significant cross contamination occurs.   The entire
solid support is then reacted with detection reagents
10   if needed and analyzed using standard colorimetric,
radioactive or fluorescent detection means.   All
processing and detection steps are performed
simultaneously to all of the microarrays on the solid
support ensuring uniform assay conditions for all of
15   the microarrays on the solid support.

       B.   Glass-Slide Polynucleotide Array
       Fig. 5 shows a substrate 136 formed according to
another aspect of the invention, and intended for use
20   in detecting binding of labeled polynucleotides to one
or more of a plurality distinct polynucleotides.   The
substrate includes a glass substrate 138 having formed
on its surface, a coating of a polycationic polymer,
preferably a cationic polypeptide, such as polylysine
25   or polyarginine.   Formed on the polycationic coating is
a microarray 140 of distinct polynucleotides, each
localized at known selected array regions, such as
regions 142.
       The slide is coated by placing a uniform-thickness
30   film of a polycationic polymer, e.g., poly-l-lysine, on
the surface of a slide and drying the film to form a
dried coating.   The amount of polycationic polymer
added is sufficient to form at least a monolayer of
polymers on the glass surface.   The polymer film is
35   bound to surface via electrostatic binding between

28

negative silyl-OH groups on the surface and charged amine groups in the polymers. Poly-1-lysine coated glass slides may be obtained commercially, e.g., from Sigma Chemical Co. (St. Louis, MO).

5   To form the microarray, defined volumes of distinct polynucleotides are deposited on the polymer-coated slide, as described in Section II. According to an important feature of the substrate, the deposited polynucleotides remain bound to the coated slide

10   surface non-covalently when an aqueous DNA sample is applied to the substrate under conditions which allow hybridization of reporter-labeled polynucleotides in the sample to complementary-sequence (single-stranded) polynucleotides in the substrate array. The method is

15   illustrated in Examples 1 and 2.

To illustrate this feature, a substrate of the type just described, but having an array of same-sequence polynucleotides, was mixed with fluorescent-labeled complementary DNA under hybridization

20   conditions. After washing to remove non-hybridized material, the substrate was examined by low-power fluorescence microscopy. The array can be visualized by the relatively uniform labeling pattern of the array regions.

25   In a preferred embodiment, each microarray contains at least 10^3 distinct polynucleotide or polypeptide biopolymers per surface area of less than about 1 cm^2. In the embodiment shown in Fig. 5, the microarray contains 400 regions in an area of about 16

30   mm^2, or 2.5×10^3 regions/cm^2. Also in a preferred embodiment, the polynucleotides in the each microarray region are present in a defined amount between about 0.1 femtomoles and 100 nanomoles in the case of polynucleotides. As above, the ability to form high-

29

density arrays of this type, where each region is
f rmed of a well-defined amount of deposited material,
can be achieved in accordance with the microarray-
forming method described in Section II.

5   Also in a preferred embodiments, the
polynucleotides have lengths of at least about 50 bp,
i.e., substantially longer than oligonucleotides which
can be formed in high-density arrays by various in situ
synthesis schemes.

10

V.   Utility

       Microarrays of immobilized nucleic acid sequences
prepared in accordance with the invention can be used
for large scale hybridization assays in numerous
15   genetic applications, including genetic and physical
mapping of genomes, monitoring of gene expression, DNA
sequencing, genetic diagnosis, genotyping of organisms,
and distribution of DNA reagents to researchers.

       For gene mapping, a gene or a cloned DNA fragment
20   is hybridized to an ordered array of DNA fragments, and
the identity of the DNA elements applied to the array
is unambiguously established by the pixel or pattern of
pixels of the array that are detected. One application
of such arrays for creating a genetic map is described
25   by Nelson, et al. (1993). In constructing physical
maps of the genome, arrays of immobilized cloned DNA
fragments are hybridized with other cloned DNA
fragments to establish whether the cloned fragments in
the probe mixture overlap and are therefore contiguous
30   to the immobilized clones on the array. For example,
Lehrach, et al., describe such a process.

       The arrays of immobilized DNA fragments may also
be used for genetic diagnostics. To illustrate, an
array containing multiple forms of a mutated gene or
35   genes can be probed with a labeled mixture of a

30

patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNA's and RNA's can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. & An example of this approach is illustrated in Examples 2, described with respect to Fig. 8.

By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm × 2.2 cm microarrays fabricated on a single 12 cm × 18 cm

31

sheet of plastic-backed nitrocellulose where each
microarray could contain, for example, 100 DNA
fragments representing all known mutations of a given
gene. The region of interest from each of the DNA
5   samples from 96 patients could be amplified, labeled,
and hybridized to the 96 individual arrays with each
assay performed in 100 microliters of hybridization
solution. The approximately 1 thick silicone rubber
barrier elements between individual arrays prevent
10   cross contamination of the patient samples by sealing
the pores of the nitrocellulose and by acting as a
physical barrier between each microarray. The solid
support containing all 96 microarrays assayed with the
96 patient samples is incubated, rinsed, detected and
15   analyzed as a single sheet of material using standard
radioactive, fluorescent, or colorimetric detection
means (Maniatas, et al., 1989). Previously, such a
procedure would involve the handling, processing and
tracking of 96 separate membranes in 96 separate sealed
20   chambers. By processing all 96 arrays as a single
sheet of material, significant time and cost savings
are possible.

     The assay format can be reversed where the patient
or organism's DNA is immobilized as the array elements
25   and each array is hybridized with a different mutated
allele or genetic marker. The gridded solid support
can also be used for parallel non-DNA ELISA assays.
Furthermore, the invention allows for the use of all
standard detection methods without the need to remove
30   the shallow barrier elements to carry out the detection
step.

     In addition to the genetic applications listed
above, arrays of whole cells, peptides, enzymes,
antibodies, antigens, receptors, ligands,
35   phospholipids, polymers, drug cogener preparations or

32

chemical substances can be fabricated by the means
described in this invention for large scale screening
assays in medical diagnostics, drug discovery,
molecular biology, immunology and toxicology.

5   The multi-cell substrate aspect of the invention
allows for the rapid and convenient screening of many
DNA probes against many ordered arrays of DNA
fragments. This eliminates the need to handle and
detect many individual arrays for performing mass

10   screenings for genetic research and diagnostic
applications. Numerous microarrays can be fabricated
on the same solid support and each microarray reacted
with a different DNA probe while the solid support is
processed as a single sheet of material.

15

        The following examples illustrate, but in no way
are intended to limit, the present invention.

                        Example 1
20   Genomic-Complexity Hybridization to Micro
                DNA Arrays Representing the Yeast
               Saccharomyces cerevisiae Genome with
                 Two-Color Fluorescent Detection
        The array elements were randomly amplified PCR
25   (Bohlander, et al., 1992) products using physically
mapped lambda clones of S. cerevisiae genomic DNA
templates (Riles, et al., 1993). The PCR was performed
directly on the lambda phage lysates resulting in an
amplification of both the 35 kb lambda vector and the

30   5-15 kb yeast insert sequences in the form of a uniform
distribution of PCR product between 250-1500 base pairs
in length.   The PCR product was purified using
Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ)
and concentrated by evaporation to dryness at room

35   temperature overnight.   Each of the 864 amplified

33

lambda clones was rehydrated in 15 μl of 3 × SSC in preparation for spotting onto the glass.

The micro arrays were fabricated on microscope slides which were coated with a layer of poly-l-lysine

5   (Sigma). The automated apparatus described in Section IV loaded 1 μl of the concentrated lambda clone PCR product in 3 × SSC directly from 96 well storage plates into the open capillary printing element and deposited ~5 nl of sample per slide at 380 micron spacing between

10   spots, on each of 40 slides. The process was repeated for all 864 samples and 8 control spots. After the spotting operation was complete, the slides were rehydrated in a humid chamber for 2 hours, baked in a dry 80° vacuum oven for 2 hours, rinsed to remove un-

15   absorbed DNA and then treated with succinic anhydride to reduce non-specific adsorption of the labeled hybridization probe to the poly-l-lysine coated glass surface. Immediately prior to use, the immobilized DNA on the array was denatured in distilled water at 90°

20   for 2 minutes.

For the pooled chromosome experiment, the 16 chromosomes of *Saccharomyces cerevisiae* were separated in a CHEF agarose gel apparatus (Biorad, Richmond, CA). The six largest chromosomes were isolated in one gel

25   slice and the smallest 10 chromosomes in a second gel slice. The DNA was recovered using a gel extraction kit (Qiagen, Chatsworth, CA). The two chromosome pools were randomly amplified in a manner similar to that used for the target lambda clones. Following

30   amplification, 5 micrograms of each of the amplified chromosome pools were separately random-primer labeled using Klenow polymerase (Amersham, Arlington Heights, IL) with a lissamine conjugated nucleotide analog (Dupont NEN, Boston, MA) for the pool containing the

35   six largest chromosomes, and with a fluorescein

34

conjugated nucle tide analog (BMB) for the pool
containing smallest ten chromosomes. The two pools
were mixed and concentrated using an ultrafiltration
device (Amicon, Danvers, MA).

5   Five micrograms of the hybridization probe
consisting of both chromosome pools in 7.5 μl of TE was
denatured in a boiling water bath and then snap cooled
on ice. 2.5 μl of concentrated hybridization solution
(5 × SSC and 0.1% SDS) was added and all 10 μl.

10   transferred to the array surface, covered with a cover
slip, placed in a custom-built single-slide humidity
chamber and incubated at 60° for 12 hours. The slides
were then rinsed at room temperature in 0.1 × SSC and
0.1%SDS for 5 minutes, cover slipped and scanned.

15   A custom built laser fluorescent scanner was used
to detect the two-color hybridization signals from the
1.8 × 1.8 cm array at 20 micron resolution. The
scanned image was gridded and analyzed using custom
image analysis software. After correcting for optical

20   crosstalk between the fluorophores due to their
overlapping emission spectra, the red and green
hybridization values for each clone on the array were
correlated to the known physical map position of the
clone resulting in a computer-generated color karyotype

25   of the yeast genome.
           Figure 6 shows the hybridization pattern of the
two chromosome pools. A red signal indicates that the
lambda clone on the array surface contains a cloned
genomic DNA segment from one of the largest six yeast

30   chromosomes. A green signal indicates that the lambda
clone insert comes from one of the smallest ten yeast
chromosomes. Orange signals indicate repetitive
sequences which cross hybridized to both chromosome
pools. Control spots on the array confirm that the

35   hybridization is specific and reproducible.

35

The physical map locations of the genomic DNA
fragments contained in each of the clones used as array
elements have been previously determined by Olson and
co-workers (Riles, et al.) allowing for the automatic
5   generation of the color karyotype shown in Figure 7.
The color of a chromosomal section on the karyotype
corresponds to the color of the array element
containing the clone from that section. The black
regions of the karyotype represent false negative dark
10   spots on the array (10%) or regions of the genome not
covered by the Olson clone library (90%). Note that
the largest six chromosomes are mainly red while the
smallest ten chromosomes are mainly green matching the
original CHEF gel isolation of the hybridization probe.
15   Areas of the red chromosomes containing green spots and
vice-versa are probably due to spurious sample tracking
errors in the formation of the original library and in
the amplification and spotting procedures.

The yeast genome arrays have also been probed with
20   individual clones or pools of clones that are
fluorescently labeled for physical mapping purposes.
The hybridization signals of these clones to the array
were translated into a position on the physical map of
yeast.
25

Example 2
Total cDNA Hybridized to Micro Arrays of
cDNA Clones with Two-Color
Fluorescent Detection

30   24 clones containing cDNA inserts from the plant
Arabidopsis were amplified using PCR. Salt was added
to the purified PCR products to a final concentration
of 3 × SSC. The cDNA clones were spotted on poly-l-
lysine coated microscope slides in a manner similar to
35   Example 1. Among the cDNA clones was a clone

36

representing a transcription factor HAT 4, which had
previously been used to create a transgenic line of the
plant *Arabidopsis*, in which this gene is present at ten
times the level found in wild-type *Arabidopsis* (Schena,
5   *et al.*, 1992).

Total poly-A mRNA from wild type *Arabidopsis* was
isolated using standard methods (Maniatis, *et al.*,
1989) and reverse transcribed into total cDNA, using
fluorescein nucleotide analog to label the cDNA product
10   (green fluorescence).   A similar procedure was
performed with the transgenic line of *Arabidopsis* where
the transcription factor HAT4 was inserted into the
genome using standard gene transfer protocols.   cDNA
copies of mRNA from the transgenic plant are labeled
15   with a lissamine nucleotide analog (red fluorescence).
Two micrograms of the cDNA products from each type of
plant were pooled together and hybridized to the cDNA
clone array in a 10 microliter hybridization reaction
in a manner similar to Example 1.   Rinsing and
20   detection of hybridization was also performed in a
manner similar to Example 1. Fig. 8 show the resulting
hybridization pattern of the array.

Genes equally expressed in wild type and the
transgenic *Arabidopsis* appeared yellow due to equal
25   contributions of the green and red fluorescence to the
final signal. The dots are different intensities of
yellow indicating various levels of gene expression.
The cDNA clone representing the transcription factor
HAT4, expressed in the transgenic line of *Arabidopsis*
30   but not detectably expressed in wild type *Arabidopsis*,
appears as a red dot (with the arrow pointing to it),
indicating the preferential expression of the
transcription factor in the red-labeled transgenic
Arabidopsis and the relative lack of expression of the

37

transcription factor in the green-labeled wild type
Arabidopsis.

An advantage of the microarray hybridization
format for gene expression studies is the high partial
5   concentration of each cDNA species achievable in the 10
microliter hybridization reaction. This high partial
concentration allows for detection of rare transcripts
without the need for PCR amplification of the
hybridization probe which may bias the true genetic
10   representation of each discrete cDNA species.

Gene expression studies such as these can be used
for genomics research to discover which genes are
expressed in which cell types, disease states,
development states or environmental conditions. Gene
15   expression studies can also be used for diagnosis of
disease by empirically correlating gene expression
patterns to disease states.

Example 3
20
Multiplexed Colorimetric Hybridization on
a Gridded Solid Support

A sheet of plastic-backed nitrocellulose was
gridded with barrier elements made from silicone rubber
according to the description in Section IV-A. The
25   sheet was soaked in 10 × SSC and allowed to dry. As
shown in Fig. 12, 192 M13 clones each with a different
yeast inserts were arrayed 400 microns apart in four
quadrants of the solid support using the automated
device described in Section III. The bottom left
30   quadrant served as a negative control for hybridization
while each of the other three quadrants was hybridized
simultaneously with a different oligonucleotide using
the open-face hybridization technology described in
Section IV-A. The first two and last four elements of

38

each array are positive controls for the colorimetric
detection step.

    The oligonucleotides were labeled with fluorescein
which was detected using an anti-fluorescein antibody
5   conjugated to alkaline phosphatase that precipitated an
NBT/BCIP dye on the solid support (Amersham). Perfect
matches between the labeled oligos and the M13 clones
resulted in dark spots visible to the naked eye and
detected using an optical scanner (HP ScanJet II)
10 attached to a personal computer. The hybridization
patterns are different in every quadrant indicating
that each oligo found several unique M13 clones from
among the 192 with a perfect sequence match. Note that
the open capillary printing tip leaves detectable
15 dimples on the nitrocellulose which can be used to
automatically align and analyze the images.

    Although the invention has been described with
respect to specific embodiments and methods, it will be
20 clear that various changes and modification may be made
without departing from the invention.

39

IT IS CLAIMED:

1.   A method of forming a microarray of analyte-
assay regions on a solid support, where each region in
5   the array has a known amount of a selected, analyte-
specific reagent, said method comprising, ·
     (a) loading a solution of a selected analyte-
specific reagent in a reagent-dispensing device having
an elongate capillary channel (i) formed by spaced-
10   apart, coextensive elongate members, (ii) adapted to
hold a quantity of the reagent solution and (iii)
having a tip region at which aqueous solution in the
channel forms a meniscus,
     (b) tapping the tip of the dispensing device
15   against a solid support at a defined position on the
surface, with an impulse effective to break the
meniscus in the capillary channel and deposit a
selected volume of solution on the surface, and
     (c) repeating steps (a) and (b) until said array
20   is formed.

2.   The method of claim 1, wherein said tapping is
carried out with an impulse effective to deposit a
selected volume in the volume range between 0.01 to 100
25   nl.

3.   The method of claim 1, wherein said channel is
formed by a pair of spaced-apart tapered elements.

30   4.   The method of claim 1, for forming a plurality
of such arrays, wherein step (b) is applied to a
selected position on each of a plurality of solid
supports at each repeat cycle proceeding step (c).

40

5. The method of claim 1, which further includes, after performing steps (a) and (b) at least one time, reloading the reagent-dispensing device with a new reagent solution by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

6. Automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent, said apparatus comprising

(a) a holder for holding, at known positions, a plurality of planar supports,

(b) a reagent dispensing device having an open capillary channel (i) formed by spaced-apart, coextensive elongate members (ii) adapted to hold a quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,

(c) positioning means for positioning the dispensing device at a selected array position with respect to a support in said holder,

(d) dispensing means for moving the device into tapping engagement against a support with a selected impulse, when the device is positioned at a defined array position with respect to that support, with an impulse effective to break the meniscus of liquid in the capillary channel and deposit a selected volume of solution on the surface, and

(e) control means for controlling said positioning and dispensing means.

41

7. The apparatus of claim 6, wherein said dispensing means is effective to move said dispensing device against a support with an impulse effective to deposit a selected volume in the volume range between
5 0.01 to 100 nl.

8. The apparatus of claim 6, wherein said channel is formed by a pair of spaced-apart tapered elements.

10 9. The apparatus of claim 6, wherein the control means operates to (i) place the dispensing device at a loading station, (ii) move the capillary channel in the device into a selected reagent at the loading station, to load the dispensing device with the reagent, and
15 (iii) dispense the reagent at a defined array position on each of the supports on said holder.

10. The apparatus of claim 6, wherein the control device further operates, at the end of a dispensing
20 cycle, to wash the dispensing device by (i) placing the dispensing device at a washing station, (ii) moving the capillary channel in the device into a wash fluid, to load the dispensing device with the fluid, and (iii) remove the wash fluid prior to loading the dispensing
25 device with a fresh selected reagent.

11. The apparatus of claim 6, wherein said device is one of a plurality of such devices which are carried on the arm for dispensing different analyte assay
30 reagents at selected spaced array positions.

12. A substrate with a surface having a microarray of at least 10^3 distinct polynucleotide or polypeptide biopolymers per 1 cm^2 surface area, each

42

distinct biopolymer sample (i) being disposed at a
separate, defined position in said array, (ii) having a
length of at least 50 subunits, and (iii) being present
in a defined amount between about 0.1 femtomole and 100
5   nanomoles.

13. The substrate of claim 12, wherein said
surface is glass slide coated with polylysine, and said
biopolymers are polynucleotides.
10

14. The substrate of claim 12, wherein said
substrate has a water-impermeable backing, a water-
permeable film formed on the backing, and a grid formed
on the film, where said grid (i) is composed of
15   intersecting water-impervious grid elements extending
from said backing to positions raised above the surface
of said film, and (ii) partitions the film into a
plurality of water-impervious cells, where each cell
contains such a biopolymer array.
20

15. A substrate with a surface array of sample-
receiving cells, comprising
        a water-impermeable backing,
        a water-permeable film formed on the backing, and
25   a grid formed on the film, said grid being composed of
intersecting water-impervious grid elements extending
from said backing to positions raised above the surface
of said film.

30   16. The substrate of claim 15, wherein the cells
of the array each contain an array of biopolymers.

17. A substrate for use in detecting binding of
labeled biopolymers to one or more of a plurality
35 distinct polynucleotides, comprising

43

a non-porous, glass substrate,

a coating of a cationic polymer on said substrate, and

5   an array of distinct polynucleotides to said coating, where each biopolymer is disposed at a separate, defined position in a surface array of biopolymers.

18.   A method of detecting differential expression
10   of each of a plurality of genes in a first cell type with respect to expression of the same genes in a second cell types, said method comprising

producing fluorescence-labeled cDNA's from mRNA's isolated from the two cells types, where the cDNA's
15   from the first and second cells are labeled with first and second different fluorescent reporters,

adding a mixture of the labeled cDNA's from the two cell types to an array of polynucleotides representing a plurality of known genes derived from
20   the two cell types, under conditions that result in hybridization of the cDNA's to complementary-sequence polynucleotides in the array; and

examining the array by fluorescence under fluorescence excitation conditions in which (i)
25   polynucleotides in the array that are hybridized predominantly to cDNA's derived from one of the first and second cell types give a distinct first or second fluorescence emission color, respectively, and (ii) polynucleotides in the array that are hybridized to
30   substantially equal numbers of cDNA's derived from the first and second cell types give a distinct combined fluorescence emission color, respectively,

wherein the relative expression of known genes in the two cell types can be determined by the observed
35   fluorescence emission color of each spot.

44

19. The method of claim 18, wherein the array of
polynucleotides is formed on a substrate with a surface
having an array of at least 10^2 distinct polynucleotide
or polypeptide biopolymers in a surface area of less
5 than about 1 cm^2, each distinct biopolymer (i) being
disposed at a separate, defined position in said array,
(ii) having a length of at least 50 subunits, and (iii)
being present in a defined amount between about .1
femtomole and 100 nmoles.

10

20. The method of claim 19, wherein said surface
is a glass slide coated with polylysine, and said
biopolymers are polynucleotides non-covalently bound to
said polylysine.

15

Fig. 1

Fig. 2A

Fig. 2B

Fig. 2C

Fig. 3

Fig. 4

4 millimeters

Fig. 5

Fig. 6

Fig. 7

Fig. 8

Fig. 9

Fig. 10

Fig. 11

Fig. 12

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :G01N 33/543, 33/68

US CL :435/6; 436/518

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 422/57; 435/4,6,973; 436/518,524,527,531,805,809

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A,P	US, A, 5,338,688 (DEEG ET AL) 16 August 1994, see entire document	1-17
A	US, A, 5,204,268 (MATSUMOTO) 20 April 1993, see entire document.	6-11
A	US, A, 4,071,315 (CHATEAU) 31 January 1978, see entire document.	12-17
A	US, A, 5,100,777 (CHANG) 31 March 1992, see entire document.	12-17
A	US, A, 5,200,312 (OPRANDY) 06 April 1993, see entire document.	12-17

☐ Further documents are listed in the continuation of Box C.   ☐ See patent family annex.

*	Special categories of cited documents:
"A"	document defining the general state of the art which is not considered to be of particular relevance
"E"	earlier document published on or after the international filing date
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O"	document referring to an oral disclosure, use, exhibition or other means
"P"	document published prior to the international filing date but later than the priority date claimed

"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&"	document member of the same patent family

Date of the actual completion of the international search

15 SEPTEMBER 1995

Date of mailing of the international search report

06 OCT 1995

Name and mailing address of the ISA/US
Commissioner of Patents and Trademarks
Box PCT
Washington, D.C. 20231

Facsimile No.   (703) 305-3230

Authorized officer

CHRISTOPHER CHIN

Telephone No.   (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

Renu A. Heller*†, Mark Schena*, Andrew Chai*, Dari Shalon‡, Tod Bedilion‡, James Gilmore‡, David E. Woolley§, and Ronald W. Davis*

*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; ‡Synteni, Palo Alto, CA 94306; and §Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

ABSTRACT   cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synoviocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Groα and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per cm². These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endoproteinases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

METHODS

Microarray Design, Development, and Preparation. Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3' end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

Abbreviations: RA, rheumatoid arthritis; MMP, matrix-degrading metalloproteinase; IBD, inflammatory bowel disease; LPS, lipopolysaccharide; PMA, phorbol 12-myristate 13-acetate; TNF-α, tumor necrosis factor α; IL, interleukin; TGF-β, transforming growth factor β; GCSF, granulocyte colony-stimulating factor; MIP, macrophage inflammatory protein; MIF, migration inhibitory factor; HME, human matrix metallo-elastase; RANTES, regulated upon activation, normal T cell expressed and secreted; Gel, gelatinase; VCAM, vascular cell adhesion molecule; ICE, IL-1 converting enzyme; PUMP, putative metalloproteinase; MnSOD, manganese superoxide dismutase; TIMP, tissue inhibitor of metalloproteinase; MCP, macrophage chemotactic protein.
†To whom reprint requests should be sent at the present address: Roche Bioscience, S3–1, 3401 Hillview Avenue, Palo Alto, CA 94304.

Biochemistry: Heller *et al.*

Proc. Natl. Acad. Sci. USA 94 (1997) 2151

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 µl of 3× standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

Tissue Specimens. Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

Cultured Cells. The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 µg/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor α (TNF-α) at 50 ng/ml, interleukin (IL)-1β at 30 ng/ml, or transforming growth factor-β (TGF-β) at 100 ng/ml is described in the figure legends.

Fluorescent Probe, Hybridization, and Scanning. Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 µl of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 µl of hybridization buffer (5× SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of β-actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

RESULTS

Ninety-Six-Gene Microarray Design. The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as β-actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

Defining Microarray Assay Conditions. Different lengths and concentrations of target DNA were tested by arraying PCR-

	1	2	3	4	5	6	7	8	9	10	11	12
A	BLANK	BLANK	HAT1 / HAT1	HAT1 / HAT1	HAT4 / HAT4	HAT4 / HAT4	HAT22 / HAT22	HAT22 / HAT22	YES23 / YES23	YES23 / YES23	BACTIN / β-actin	G3PDH / G3PDH
B	IL1A / IL-1α	IL1B / IL-1β	IL1RA / IL-1RA	IL2 / IL-2	IL3 / IL-3	IL4 / IL-4	IL6 / IL-6	IL6R / IL-6R	IL7 / IL-7	CFOS / c-fos	CJUN / c-jun	RFRA1 / Rat Fra-1
C	IL8 / IL-8	IL9 / IL-9	IL10 / IL-10	ICE / ICE	IFNG / IFNγ	GCSF / G-CSF	MCSF / M-CSF	GMCSF / GM-CSF	TNFB.1 / TNFβ	CREL / c-rel	NFKB50 / NFκBp50	NFKB65.1 / NFκBp65
D	TNFA.1 / TNFα	TNFA.2 / TNFα	TNFA.3 / TNFα	TNFA.4 / TNFα	TNFA.5 / TNFα	TNFRI.1 / TNFrI	TNFRI.2 / TNFrI	TNFRII.1 / TNFrII	TNFRII.2 / TNFrII	NFKB65.2 / NFκBp65	IKB / IκB	CREB2 / CREB2
E	STR1 / Strom-1	STR2-3 / Strom-2	STR3 / Strom-3	COL1 / Coll-1	COL1-3 / Coll-1.3	COL2.1 / Coll-2	COL2.2 / Coll-2	COL3 / Coll-3	COX1 / Cox-1	COX2 / Cox-2	12LO / 12-LO	15LO / 15-LO
F	GELA.1 / Gel-A	GELB / Gel-B	HME / Elastase	MTMMP / MT-MMP	PUMP1 / Matrilysin	TIMP1 / TIMP-1	TIMP2 / TIMP-2	TIMP3 / TIMP-3	ICAM1 / ICAM-1	VCAM / VCAM	5LO.1 / 5-LO	CPLA2.2 / cPLA2
G	EGF / EGF	FGFA / FGF-acidic	FGFB / FGF-basic	IGFI / IGF-I	IGFII / IGF-II	TGFA / TGFα	TGFB / TGFβ	PDGFB / PDGFβ	CALCTN / Calcitonin	GH1 / GH-1	GRO / GRO1α	GCR / GR
H	MCP1.1 / MCP-1	MCP1.1 / MCP-1	MIP1A / MIP-1α	MIP1B / MIP-1β	MIF / MIF	RANTES / RANTES	INOS / INOS	LDLR / LDLR	ALU.1 / IL-10	ALU.2 / TNFRp70	ALU.3 / IL-10	POLYA / LDLR

▨ *A. thaliana* controls	□ Cytokines and related genes	▨ Chemokines
▨ Human controls	Transcription factors and related genes	▨ Growth factors and related genes
	□ MMP's and related genes	▨ Other genes

FIG. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 μg/μl or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 μg/μl sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ≈300 bp, arrayed at a concentration of 1 μg/μl. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3' proximal regions, because the signal was reduced if the target fragment was biased toward the 5' end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3' proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metal-

loproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

Monitoring Differential Expression in Cultured Cell Lines. In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-
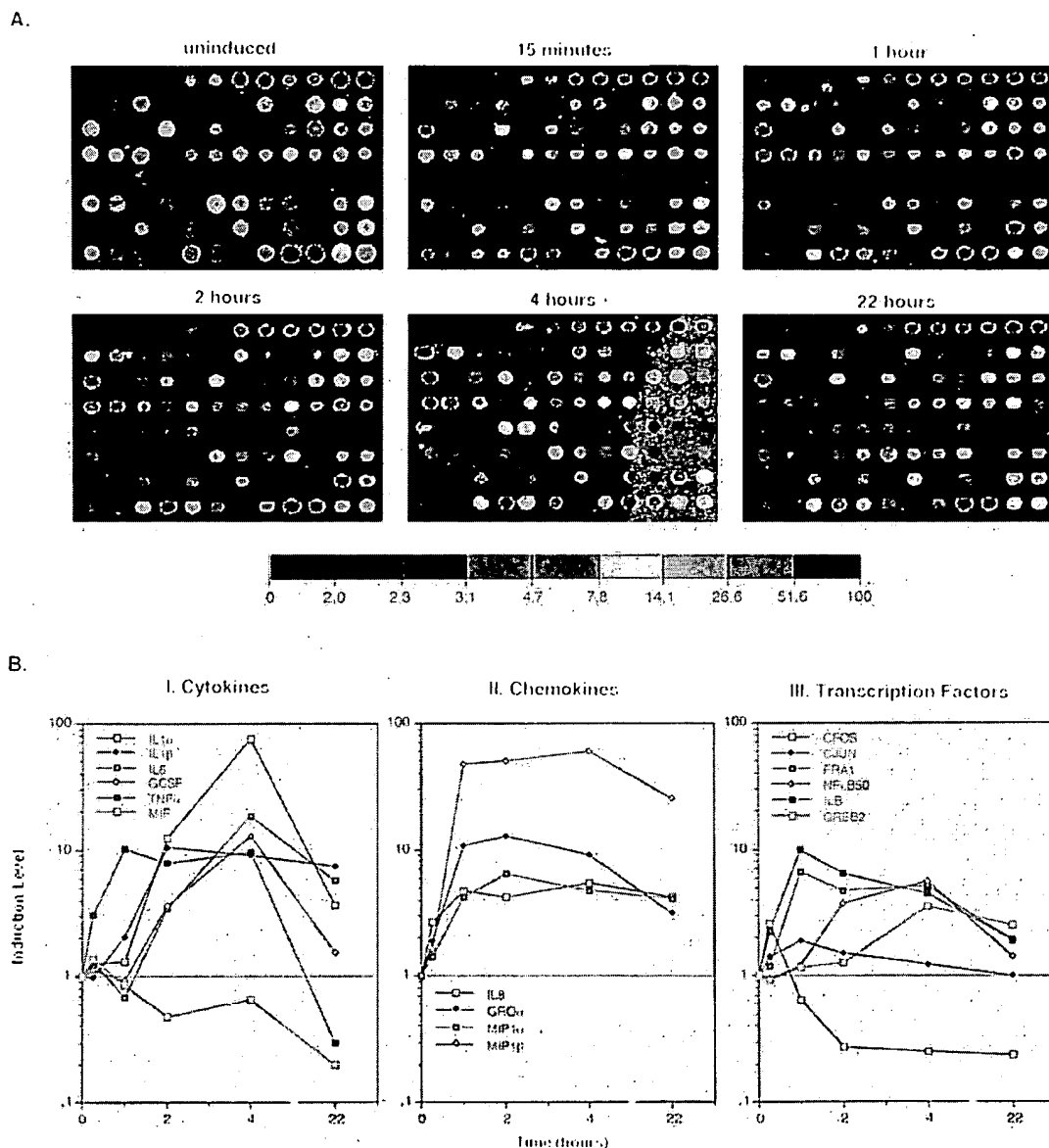

FIG. 2.   Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (*A*) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (*B I–III*) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

ically these cells, when triggered by an immunogen, produce the proinflammatroy cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 2A). Many other cytokine genes were also transiently activated, such as IL-1α and -β, IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1β, more so than MIP-1α, and Groα or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were c-*fos*, *fra-1*, c-*jun*, NF-κBp50, and IκB, with c-*rel* expression observed even in the uninduced state (Fig. 2B). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-

type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1), -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 3A). HME induction was estimated to be ≈50-fold and was greater than any of the other MMPs examined (Fig. 3B). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1β was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

Quantitation of differential gene expression (Figs. 2B and 3B) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1β, IL-8, MIP-1β, TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue. Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of c-*jun*, GCSF, IL-3, TNF-β, MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, GelA, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3′ end of the gene gave no
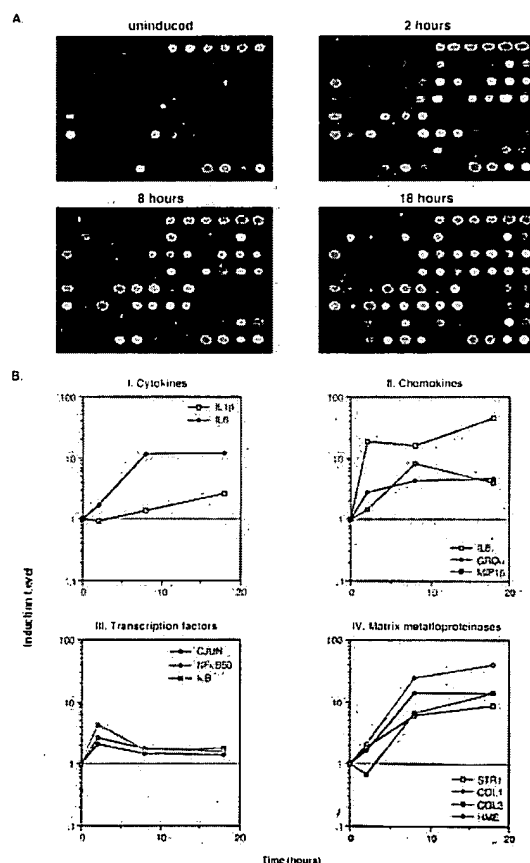

FIG. 3. Time course for IL-1β and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (A) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (B I–IV) Relative levels of selected genes at different time points compared with time zero.
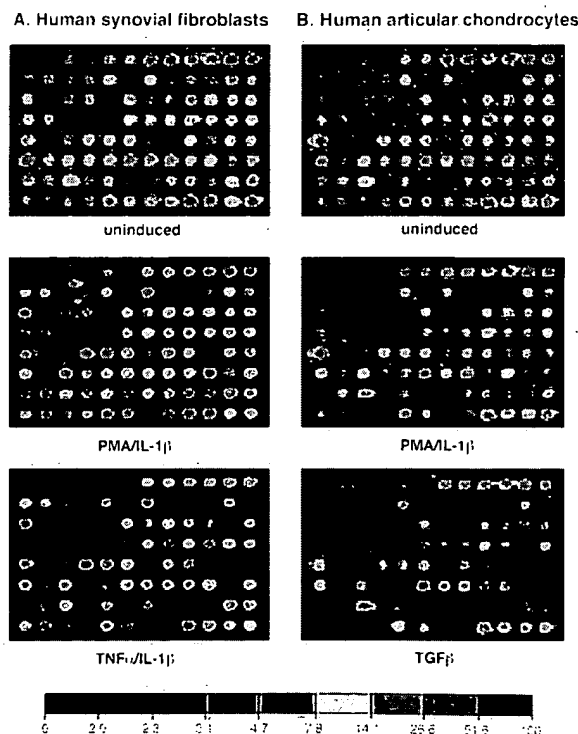

FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1β, or TNF and IL-1β, or TGF-β for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1α, and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF-β had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in
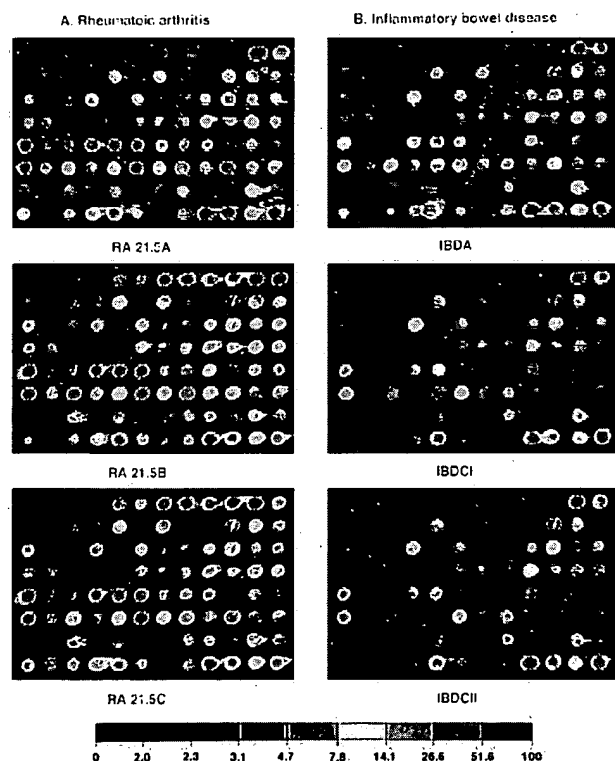

FIG. 5. Expression profiles of RA tissue (A) and IBD tissue (B). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Groα, and MMP expression was more pronounced than in the others.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs were sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1α, MIP-1β, IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1α and IL-β followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarachy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were

the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF-β was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Groα, MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor β was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Groα. With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1β processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synoviocytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Groα, is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It downregulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synoviocytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1β, the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new

targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
2. Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* **6**, 639–645.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619.
4. Feldmann, M., Brennan F. M. & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* **85**, 307–310.
5. Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410–460.
6. Lotz, M. F., Blanco, J., Von Kempis, J., Dudler, J., Maier, R., Villiger P. M. & Geng, Y. (1995) *J. Rheumatol.* **22**, Supplement 43, 104–108.
7. Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A. &. Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* **4**, 197–250.
8. Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A. & Reithmuller, G. (1988) *Int. J. Cancer* **41**, 456–461.
9. Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P. & Heller, R. A. (1996) *J. Biol. Chem.* **271**, 23577–23581.
10. Gadher, S. J. & Woolley, D. E. (1987) *Rheumatol. Int.* **7**, 13–22.
11. Harris, E. D., Jr. (1990) *New Engl. J. Med.* **322**, 1277–1289.
12. Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B. (Saunders, Philadelphia), 5th Ed. pp. 5001–5047.
13. Alvaro-Garcia, J. M., Zvaifler, Nathan J., Brown, C. B., Kaushansky, K. & Firestein, Gary S. (1991) *J. Immunol.* **146**, 3365–3371.
14. Firestein, G. S., Alvaro-Grarcia, J. M. & Maki, R. (1990) *J. Immunol.* **144**, 3347–3352.
15. Pradines-Figueres, A. & Raetz, C. R. H. (1992) *J. Biol. Chem.* **267**, 23261–23268.
16. Shapiro, S. D., Kobayashi, D. L. & Ley, T. J. (1993) *J. Biol. Chem.* **208**, 23824–23829.
17. Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J. & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3042–3946.
18. Cerreti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K. & Black, R. A. (1992) *Science* **256**, 97–100.
19. Miura, M., Zhu, H., Rotello, R., Hartweig, E. A. & Yuan, J. (1993) *Cell* **75**, 653–660.
20. Arai, K., Lee, F., Miyajima, A., Shoichiro, M., Arai, N. & Takashi, Y. (1990) *Annu. Rev. Biochem.* **59**, 783–836.
21. Geiser, T., Dewald, B., Ehrengruber, M. U., Lewis, I. C. & Baggiolini, M. (1993) *J. Biol. Chem.* **268**, 15419–15424.
22. Unemori, E. N., Amento, E. P., Bauer, E. A. & Horuk, R. (1993) *J. Biol. Chem.* **268**, 1338–1342.
23. Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N. & Fish, E. N. (1995) *Clin. Exp. Immunol.* **101**, 398–407.
24. Roeser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A. & Worwood, M. (Academic, New York), Vol. 2, pp. 605–640.
25. Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V. & Torti, F. M. (1995) *J. Biol. Chem.* **270**, 15285–15293.

PCT

WORLD INTELLECTUAL PROPERTY ORGANIZATION
International Bureau

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : **C12Q 1/68, C07H 21/04**	**A1**	(11) International Publication Number: **WO 97/13877** (43) International Publication Date: 17 April 1997 (17.04.97)

(21) International Application Number: PCT/US96/16342

(22) International Filing Date: 11 October 1996 (11.10.96)

(30) Priority Data:
PCT/US95/12791 12 October 1995 (12.10.95) WO
*(34) Countries for which the regional or
international application was filed:* US et al.
PCT/US96/09513 6 June 1996 (06.06.96) WO
*(34) Countries for which the regional or
international application was filed:* US et al.

(60) Parent Application or Grant
(63) Related by Continuation
US Not furnished (CIP)
Filed on Not furnished

(71) Applicant *(for all designated States except US):* LYNX THER-
APEUTICS, INC. [US/US]; 3832 Bay Center Place, Hay-
ward, CA 94545 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only):* MARTIN, David, W.
[US/US]; Lynx Therapeutics, Inc., 3832 Bay Center Place,
Hayward, CA 94545 (US).

(74) Agent: POWERS, Vincent, M.; Dehlinger & Associates, Post
Office Box 60850, Palo Alto, CA 94306-0850 (US).

(81) Designated States: AU, CA, CZ, EE, FI, HU, JP, KR, LT, LV,
NO, NZ, PL, RU, SG, US, European patent (AT, BE, CH,
DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT,
SE).

Published
With international search report.
*Before the expiration of the time limit for amending the
claims and to be republished in the event of the receipt of
amendments.*

(54) Title: MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION

(57) Abstract

A method is provided for assessing the toxicity of a compound in a test organism by measuring gene expression profiles of selected
tissues. Gene expression profiles are measured by massively parallel signature sequencing of cDNA libraries constructed from mRNA
extracted from the selected tissues. Gene expression profiles provide extensive information on the effects of administering a compound to
a test organism in both acute toxicity tests and in prolonged and chronic toxicity tests.

FOR THE PURPOSES OF INFORMATION ONLY

Codes used to identify States party to the PCT on the front pages of pamphlets publishing international applications under the PCT.

AM	Armenia	GB	United Kingdom	MW	Malawi
AT	Austria	GE	Georgia	MX	Mexico
AU	Australia	GN	Guinea	NE	Niger
BB	Barbados	GR	Greece	NL	Netherlands
BE	Belgium	HU	Hungary	NO	Norway
BF	Burkina Faso	IE	Ireland	NZ	New Zealand
BG	Bulgaria	IT	Italy	PL	Poland
BJ	Benin	JP	Japan	PT	Portugal
BR	Brazil	KE	Kenya	RO	Romania
BY	Belarus	KG	Kyrgystan	RU	Russian Federation
CA	Canada	KP	Democratic People's Republic	SD	Sudan
CF	Central African Republic		of Korea	SE	Sweden
CG	Congo	KR	Republic of Korea	SG	Singapore
CH	Switzerland	KZ	Kazakhstan	SI	Slovenia
CI	Côte d'Ivoire	LI	Liechtenstein	SK	Slovakia
CM	Cameroon	LK	Sri Lanka	SN	Senegal
CN	China	LR	Liberia	SZ	Swaziland
CS	Czechoslovakia	LT	Lithuania	TD	Chad
CZ	Czech Republic	LU	Luxembourg	TG	Togo
DE	Germany	LV	Latvia	TJ	Tajikistan
DK	Denmark	MC	Monaco	TT	Trinidad and Tobago
EE	Estonia	MD	Republic of Moldova	UA	Ukraine
ES	Spain	MG	Madagascar	UG	Uganda
FI	Finland	ML	Mali	US	United States of America
FR	France	MN	Mongolia	UZ	Uzbekistan
GA	Gabon	MR	Mauritania	VN	Viet Nam

MEASUREMENT OF GENE EXPRESSION PROFILES
IN TOXICITY DETERMINATION

Field of the Invention

The invention relates generally to methods for detecting and monitoring phenotypic changes in in vitro and in vivo systems for assessing and/or determining the toxicity of chemical compounds, and more particularly, the invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro and in vivo systems for determining the toxicity of drug candidates.

BACKGROUND

The ability to rapidly and conveniently assess the toxicity of new compounds is extremely important. Thousands of new compounds are synthesized every year, and many are introduced to the environment through the development of new commercial products and processes, often with little knowledge of their short term and long term health effects. In the development of new drugs, the cost of assessing the safety and efficacy of candidate compounds is becoming astronomical: It is .estimated that the pharmaceutical industry spends an average of about 300 million dollars to bring a new pharmaceutical compound to market. e.g. Biotechnology. 13: 226-228 (1995). A large fraction of these costs are due to the failure of candidate compounds in the later stages of the developmental process. That is. as the assessment of a candidate drug progresses from the identification of a compound as a drug candidate--for example. through relatively inexpensive binding assays or in vitro screening assays, to pharmacokinetic studies. to toxicity studies. to efficacy studies in model systems. to preliminary clinical studies; and so on, the costs of the associated tests and analyses increases tremendously. Consequently, it may cost several tens of millions of dollars to determine that a once promising candidate compound possesses a side effect or cross reactivity that renders it commercially infeasible to develop further. A great challenge of pharmaceutical development is to remove from further consideration as early as possible those compounds that are likely to fail in the later stages of drug testing.

Drug development programs are clearly structured with this objective in mind; however, rapidly escalating costs have created a need to develop even more stringent and less expensive screens in the early stages to identify false leads as soon as possible. Toxicity assessment is an area where such improvements may be made. for both drug development and for assessing the environmental. health. and safety effects of new compounds in general.

Typically the toxicity of a compound is determined by administering the compound to one or more species of test animal under controlled conditions and by monitoring the effects on a wide range of parameters. The parameters include such things as blood chemistry, weight gain or loss, a variety of behavioral patterns, muscle

5   tone, body temperature, respiration rate, lethality, and the like, which collectively provide a measure of the state of health of the test animal. The degree of deviation of such parameters from their normal ranges gives a measure of the toxicity of a compound. Such tests may be designed to assess the acute, prolonged, or chronic toxicity of a compound. In general, acute tests involve administration of the test

10   chemical on one occasion. The period of observation of the test animals may be as short as a few hours, although it is usually at least 24 hours and in some cases it may be as long as a week or more. In general, prolonged tests involve administration of the test chemical on multiple occasions. The test chemical may be administered one or more times each day, irregularly as when it is incorporated in the diet, at specific

15   times such as during pregnancy, or in some cases regularly but only at weekly intervals. Also, in the prolonged test the experiment is usually conducted for not less than 90 days in the rat or mouse or a year in the dog. In contrast to the acute and prolonged types of test, the chronic toxicity tests are those in which the test chemical . is administered for a substantial portion of the lifetime of the test animal. In the case

20   of the mouse or rat, this is a period of 2 to 3 years. In the case of the dog, it is for 5 to 7 years.

Significant costs are incurred in establishing and maintaining large cohorts of test animals for such assays, especially the larger animals in chronic toxicity assays. Moreover, because of species specific effects, passing such toxicity tests does not

25   ensure that a compound is free of toxic effects when used in humans. Such tests do, however, provide a standardized set of information for judging the safety of new compounds, and they provide a database for giving preliminary assessments of related compounds. An important area for improving toxicity determination would be the identification of new observables which are predictive of the outcome of the

30   expensive and tedious animal assays.

In other medical fields, there has been significant interest in applying recent advances in biotechnology, particularly in DNA sequencing, to the identification and study of differentially expressed genes in healthy and diseased organisms, e.g. Adams et al, Science, 252: 1651-1656 (1991); Matsubara et al, Gene, 135: 265-274 (1993);

35   Rosenberg et al, International patent application, PCT/US95/01863. The objectives of such applications include increasing our knowledge of disease processes, identifying genes that play important roles in the disease process, and providing diagnostic and therapeutic approaches that exploit the expressed genes or their

products. While such approaches are attractive, those based on exhaustive, or even sampled, sequencing of expressed genes are still beset by the enormous effort required: It is estimated that 30-35 thousand different genes are expressed in a typical mammalian tissue in any given state, e.g. Ausubel et al, Editors, Current Protocols,

5   5.8.1-5.8.4 (John Wiley & Sons, New York, 1992). Determining the sequences of even a small sample of that number of gene products is a major enterprise, requiring industrial-scale resources. Thus, the routine application of massive sequencing of expressed genes is still beyond current commercial technology.

The availability of new assays for assessing the toxicity of compounds, such

10   as candidate drugs, that would provide more comprehensive and precise information about the state of health of a test animal would be highly desirable. Such additional assays would preferably be less expensive, more rapid, and more convenient than current testing procedures, and would at the same time provide enough information to make early judgments regarding the safety of new compounds.

15

Summary of the Invention

An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo .test systems.

20   Another object of the invention is to provide a database on which to base decisions concerning the toxicological properties of chemicals, particularly drug candidates.

A further object of the invention is to provide a method for analyzing gene expression patterns in selected tissues of test animals.

25   A still further object of the invention is to provide a system for identifying genes which are differentially expressed in response to exposure to a test compound.

Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals.

Another object of the invention is to identify genes whose expression is

30   predictive of deleterious toxicity.

The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel DNA sorting and sequencing methodologies that

35   permit the formation of gene expression profiles for selected tissues by determining the sequence of portions of many thousands of different polynucleotides in parallel. Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity.

The sorting methodology of the invention makes use of oligonucleotide tags that are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set
5   cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags of the invention, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements
10   provide a means of enhancing specificity of hybridization for sorting polynucleotides, such as cDNAs.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully below, this condition is achieved by employing a repertoire of tags substantially
15   greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or
20   regions. The sorted populations of polynucleotides can then be sequenced on the solid phase support by a "single-base" or "base-by-base" sequencing methodology, as described more fully below.

In one aspect, the method of the invention comprises the following steps: (a) administering the compound to a test organism; (b) extracting a population of mRNA
25   molecules from each of one or more tissues of the test organism; (c) forming a separate population of cDNA molecules from each population of mRNA molecules extracted from the one or more tissues such that each cDNA molecule of the separate populations has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set; (d) separately sampling each
30   population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached; (e) sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in
35   spatially discrete regions on one or more solid phase supports; (f) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of

the one or more tissues; and (g) correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

    An important aspect of the invention is the identification of genes whose expression is predictive of the toxicity of a compound. Once such genes are

5   identified, they may be employed in conventional assays, such as reverse transcriptase polymerase chain reaction (RT-PCR) assays for gene expression.

Brief Description of the Drawings

    Figure 1 is a flow chart representation of an algorithm for generating

10   minimally cross-hybridizing sets of oligonucleotides.

    Figure 2 diagrammatically illustrates an apparatus for carrying out polynucleotide sequencing in accordance with the invention.

Definitions

15   "Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be .selected to be either double stranded or single stranded. Thus, where triplexes are

20   formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

    The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides,

25   anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form

30   oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless

35   otherwise noted. Analogs of phosphodiester linkages include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the

art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990). or the like. with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties. reduce complexity, increase specificity, and the like.

As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides, usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide. e.g. cytosines, within the target polynucleotide "CATCGC ..." so that its sequence is represented as a binary code, e.g. "100101 ... " for "C-(not C)-(not C)-C-(not C)-C ... " and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of molecule present in the population.

5   As used herein, the terms "gene expression profile," and "gene expression pattern" which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. Generally, the portions of sequence are sufficiently long to uniquely identify the cDNA from which the portion arose. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences

10   determined in a gene expression profile is at least ten thousand.

As used herein, "test organism" means any in vitro or in vivo system which provides measureable responses to exposure to test compounds. Typically, test organisms may be mammalian cell cultures, particularly of specific tissues, such as hepatocytes, neurons, kidney cells, colony forming cells, or the like, or test organisms

15   may be whole animals, such as rats, mice, hamsters, guinea pigs, dogs, cats, rabbits, pigs, monkeys, and the like.

Detailed Description of the Invention

The invention provides a method for determining the toxicity of a compound

20   by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. The invention also provides a method of identifying toxicity markers consisting of individual genes or a group of genes that is expressed acutely and which is correlated with prolonged or chronic toxicity, or suggests that the compound will have an undesirable cross reactivity. Gene

25   expression profiles are generated by sequencing portions of cDNA molecules construction from mRNA extracted from tissues of test organisms exposed to the compound being tested. As used herein, the term "tissue" is employed with its usual medical or biological meaning, except that in reference to an in vitro test system, such as a cell culture, it simply means a sample from the culture. Gene expression profiles

30   derived from test organisms are compared to gene expression profiles derived from control organisms to determine the genes which are differentially expressed in the test organism because of exposure to the compound being tested. In both cases, the sequence information of the gene expression profiles is obtained by massively parallel signature sequencing of cDNAs, which is implemented in steps (c) through (f) of the

35   above method.

Toxicity Assessment

Procedures for designing and conducting toxicity tests in in vitro and in vivo systems is well known, and is described in many texts on the subject, such as Loomis

et al. Loomis's Esstentials of Toxicology, 4th Ed. (Academic Press, New York, 1996);
Echobichon, The Basics of Toxicity Testing (CRC Press, Boca Raton, 1992); Frazier,
editor, In Vitro Toxicity Testing (Marcel Dekker, New York, 1992); and the like.

In toxicity testing, two groups of test organisms are usually employed: one
group serves as a control and the other group receives the test compound in a single
dose (for acute toxicity tests) or a regimen of doses (for prolonged or chronic toxicity
tests). Since in most cases, the extraction of tissue as called for in the method of the
invention requires sacrificing the test animal, both the control group and the group
receiving compound must be large enough to permit removal of animals for sampling
tissues, if it is desired to observe the dynamics of gene expression through the
duration of an experiment.

In setting up a toxicity study, extensive guidance is provided in the literature
for selecting the appropriate test organism for the compound being tested, route of
administration, dose ranges, and the like. Water or physiological saline (0.9% NaCl
in water) is the solute of choice for the test compound since these solvents permit
administration by a variety of routes. When this is not possible because of solubility
limitations, it is necessary to resort to the use of vegetable oils such as corn oil or
even organic solvents, of which propylene glycol is commonly used. Whenever
-possible the use of suspension of emulsion should be avoided except for oral
administration. Regardless of the route of administration, the volume required to
administer a given dose is limited by the size of the animal that is used. It is desirable
to keep the volume of each dose uniform within and between groups of animals.
When rates or mice are used the volume administered by the oral route should not
exceed 0.005 ml per gram of animal. Even when aqueous or physiological saline
solutions are used for parenteral injection the volumes that are tolerated are limited,
although such solutions are ordinarily thought of as being innocuous. The
intravenous LD_{50} of distilled water in the mouse is approximately 0.044 ml per gram
and that of isotonic saline is 0.068 ml per gram of mouse.

When a compound is to be administered by inhalation, special techniques for
generating test atmospheres are necessary. Dose estimation becomes very
complicated. The methods usually involve aerosolization or nebulization of fluids
containing the compound. If the agent to be tested is a fluid that has an appreciable
vapor pressure, it may be administered by passing air through the solution under
controlled temperature conditions. Under these conditions, dose is estimated from the
volume of air inhaled per unit time, the temperature of the solution, and the vapor
pressure of the agent involved. Gases are metered from reservoirs. When particles of
a solution are to be administered, unless the particle size is less than about 2 μm the
particles will not reach the terminal alveolar sacs in the lungs. A variety of

apparatuses and chambers are available to perform studies for detecting effects of irritant or other toxic endpoints when they are administered by inhalation. The preferred method of administering an agent to animals is via the oral route, either by intubation or by incorporating the agent in the feed.

5   Preferably, in designing a toxicity assessment, two or more species should be employed that handle the test compound as similarly to man as possible in terms of metabolism, absorption, excretion, tissue storage, and the like. Preferably, multiple doses or regimens at different concentrations should be employed to establish a dose-response relationship with respect to toxic effects. And preferably, the route of

10   administration to the test animal should be the same as, or as similar as possible to, the route of administration of the compound to man. Effects obtained by one route of administration to test animals are not a priori applicable to effects by another route of administration to man. For example, food additives for man should be tested by admixture of the material in the diet of the test animals.

15   Acute toxicity tests consist of administering a compound to test organisms on one occasion. The purpose of such test is to determine the symptomotology consequent to administration of the compound and to determine the degree of lethality of the compound. The initial procedure is to perform a series of range-finding doses of the compound in a single species. This necessitates selection of a route of

20   administration, preparation of the compound in a form suitable for administration by the selected route, and selection of an appropriate species. Preferably, initial acute toxicity studies are performed on either rats or mice because of their low cost, their availability, and the availability of abundant toxicologic reference data on these species. Prolonged toxicity tests consist of administering a compound to test

25   organisms repeatedly, usually on a daily basis, over a period of 3 to 4 months. Two practical factors are encountered that place constraints on the design of such tests: First, the available routes of administration are limited because the route selected must be suitable for repeated administration without inducing harmful effects. And second, blood, urine, and perhaps other samples, should be taken repeatedly without

30   inducing significant harm to the test animals. Preferably, in the method of the invention the gene expression profiles are obtained in conjunction with the measurement of the traditional toxicologic parameters, such as listed in the table below:

35

Hematology	Blood Chemistry	Urine Analyses
erythrocyte count	sodium	pH
total leukocyte count	potassium	specific gravity
differential leukocyte count	chloride	total protein
hematocrit	calcium	sediment
hemoglobin	carbon dioxide	glucose
	serum glutamine-pyruvate transaminase	ketones
	serum glutamin-oxalacetic transaminase	bilirubin
	serum protein electrophoresis	
	blood sugar	
	blood urea nitrogen	
	total serum protein	
	serum albumin	
	total serum bilirubin	

5   Oligonucleotide Tags and Tag Complements

Oligonucleotide tags are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set cannot form a duplex (or triplex) with the complement of
10   any other member with less than two mismatches. Complements of oligonucleotide tags, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements provide a means of enhancing specificity of hybridization for sorting,
15   tracking, or labeling molecules, especially polynucleotides.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements may be synthesized either combinatorially or individually depending on the size of the set desired and the degree to which cross-hybridization is sought to be minimized (or stated another way, the degree to which specificity is sought to be enhanced). For
20   example, a minimally cross-hybridizing set may consist of a set of individually synthesized 10-mer sequences that differ from each other by at least 4 nucleotides. such set having a maximum size of 332 (when composed of 3 kinds of nucleotides and counted using a computer program such as disclosed in Appendix Ic). Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be

assembled combinatorially from subunits which themselves are selected from a minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing 12-mers differing from one another by at least three nucleotides may be synthesized by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers

5   that each differ from one another by three nucleotides. Such an embodiment gives a maximally sized set of 9^3, or 729, 12-mers. The number 9 is number of oligonucleotides listed by the computer program of Appendix Ia, which assumes, as with the 10-mers, that only 3 of the 4 different types of nucleotides are used. The set is described as "maximal" because the computer programs of Appendices Ia-c provide

10   the largest set for a given input (e.g. length, composition, difference in number of nucleotides between members). Additional minimally cross-hybridizing sets may be formed from subsets of such calculated sets.

Oligonucleotide tags may be single stranded and be designed for specific hybridization to single stranded tag complements by duplex formation or for specific

15   hybridization to double stranded tag complements by triplex formation. Oligonucleotide tags may also be double stranded and be designed for specific hybridization to single stranded tag complements by triplex formation.

When synthesized combinatorially, an oligonucleotide tag preferably consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9

20   nucleotides in length wherein each subunit is selected from the same minimally cross-hybridizing set. In such embodiments, the number of oligonucleotide tags available depends on the number of subunits per tag and on the length of the subunits. The number is generally much less than the number of all possible sequences the length of the tag, which for a tag n nucleotides long would be 4^n.

25   Complements of oligonucleotide tags attached to a solid phase support are used to sort polynucleotides from a mixture of polynucleotides each containing a tag. Complements of the oligonucleotide tags are synthesized on the surface of a solid phase support, such as a microscopic bead or a specific location on an array of synthesis locations on a single support, such that populations of identical sequences

30   are produced in specific regions. That is, the surface of each support, in the case of a bead, or of each region, in the case of an array, is derivatized by only one type of complement which has a particular sequence. The population of such beads or regions contains a repertoire of complements with distinct sequences. As used herein in reference to oligonucleotide tags and tag complements, the term "repertoire" means

35   the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in a particular embodiment or the corresponding set of tag complements.

The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully

below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed
5   under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions.

        The nucleotide sequences of oligonucleotides of a minimally cross-hybridizing set are conveniently enumerated by simple computer programs, such as those
10   exemplified by programs whose source codes are listed in Appendices Ia and Ib. Program minhx of Appendix Ia computes all minimally cross-hybridizing sets having 4-mer subunits composed of three kinds of nucleotides. Program tagN of Appendix Ib enumerates longer oligonucleotides of a minimally cross-hybridizing set. Similar algorithms and computer programs are readily written for listing oligonucleotides of
15   minimally cross-hybridizing sets for any embodiment of the invention. Table I below provides guidance as to the size of sets of minimally cross-hybridizing oligonucleotides for the indicated lengths and number of nucleotide differences. The above computer programs were used to generate the numbers.

20   Table I

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross-Hybridizing Set	Maximal Size of Minimally Cross-Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	3	9	6561	5.90×10^4
6	3	27	5.3×10^5	1.43×10^7
7	4	27	5.3×10^5	1.43×10^7
7	5	8	4096	3.28×10^4
8	3	190	1.30×10^9	2.48×10^{11}
8	4	62	1.48×10^7	9.16×10^8
8	5	18	1.05×10^5	1.89×10^6
9	5	39	2.31×10^6	9.02×10^7
10	5	332	1.21×10^{10}	
10	6	28	6.15×10^5	1.72×10^7
11	5	187		
18	6	≈25000		

18   12   24

For some embodiments of the invention, where extremely large repertoires of tags are not required, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized. Sets containing several hundred to several thousands, or
5   even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, e.g. as disclosed in Frank et al, U.S. patent 4,689,405; Frank et al, Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al, Anal. Biochem., 224: 110-116 (1995); Fodor et al, International application PCT/US93/04145; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994);
10   Southern et al, J. Biotechnology, 35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al, Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, oligonucleotide tags of the invention are synthesized combinatorially out of subunits between three and six nucleotides in length and
15   selected from the same minimally cross-hybridizing set. For oligonucletides in this range, the members of such sets may be enumerated by computer programs based on the algorithm of Fig. 1.

The algorithm of Fig. 1 is implemented by first defining the characteristics of the subunits of the minimally cross-hybridizing set, i.e. length, number of base
20   differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table M_n, n=1, is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit S_1 is selected and compared (120) with successive subunits S_i for i=n+1 to the end of the table. Whenever a successive subunit has the required number of mismatches to be a
25   member of the minimally cross-hybridizing set, it is saved in a new table M_{n+1} (125), that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons, M_2 will contain S_1; in the second set of comparisons, M_3 will contain S_1 and S_2; in the third set of comparisons, M_4 will contain S_1, S_2, and S_3; and so on. Similarly, comparisons in table M_j will be
30   between S_j and all successive subunits in M_j. Note that each successive table M_{n+1} is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table M_n has been compared (140) the old table is replaced by the new table M_{n+1}, and the next round of comparisons are begun. The process stops (160) when a table M_n is reached that contains no
35   successive subunits to compare to the selected subunit S_i, i.e. $M_n = M_{n+1}$.

Preferably, minimally cross-hybridizing sets comprise subunits that make approximately equivalent contributions to duplex stability as every other subunit in

the set. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989)

5   and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991);and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA

10   Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed. Clearly, the are many approaches available to one skilled in the art for designing sets of minimally cross-hybridizing subunits within the scope of the invention. For example, to minimize the affects of different base-stacking energies of terminal nucleotides when subunits are assembled, subunits may be provided that

15   have the same terminal nucleotides. In this way, when subunits are linked, the sum of the base-stacking energies of all the adjoining terminal nucleotides will be the same, thereby reducing or eliminating variability in tag melting temperatures.

A "word" of terminal nucleotides, shown in italic below, may also be added to each end of a tag so that a perfect match is always formed between it and a similar

20   terminal "word" on any other tag complement. Such an augmented tag would have the form:

W	W_1	W_2	...	W_{k-1}	W_k	W
W'	W_1'	W_2'	...	W_{k-1}'	W_k'	W'

where the primed W's indicate complements. With ends of tags always forming

25   perfectly matched duplexes, all mismatched words will be internal mismatches thereby reducing the stability of tag-complement duplexes that otherwise would have mismatched words at their ends. It is well known that duplexes with internal mismatches are significantly less stable than duplexes with the same mismatch at a terminus.

30   A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the $5' \rightarrow 3'$ exonuclease activity of a DNA polymerase. The following is an exemplary

35   minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

5   Table II

Word:	w_1	w_2	w_3	w_4
Sequence:	GATT	TGAT	TAGA	TTTG

Word:	w_5	w_6	w_7	w_8
Sequence:	GTAA	AGTA	ATGT	AAAG

In this set, each member would form a duplex having three mismatched bases with
10   the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table
III. Clearly, additional sets can be generated by substituting different groups of
nucleotides, or by using subsets of known minimally cross-hybridizing sets.

15

                               Table III
           Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG

- 15 -

Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting _oligonucleotides are capable of specific hybridization. In some embodiments, tags may comprise naturally occurring nucleotides that permit processing or manipulation by enzymes, while the corresponding tag complements may comprise non-natural nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements may be generated by subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al. International patent application PCT/US93/03418 or Lyttle et al, Biotechniques, 19: 274-280 (1995). Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J. Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of

mixtures of the activated monomers to the growing oligonucleotide during the coupling steps. Preferably, oligonucleotide tags and tag complements are synthesized on a DNA synthesizer having a number of synthesis chambers which is greater than or equal to the number of different kinds of words used in the construction of the tags.

5   That is, preferably there is a synthesis chamber corresponding to each type of word. In this embodiment, words are added nucleotide-by-nucleotide, such that if a word consists of five nucleotides there are five monomer couplings in each synthesis chamber. After a word is completely synthesized, the synthesis supports are removed from the chambers, mixed, and redistributed back to the chambers for the next cycle

10   of word addition. This latter embodiment takes advantage of the high coupling yields of monomer addition, e.g. in phosphoramidite chemistries.

Double stranded forms of tags may be made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Alternatively, double stranded tags may be formed by first synthesizing a

15   single stranded repertoire linked to a known oligonucleotide sequence that serves as a primer binding site. The second strand is then synthesized by combining the single stranded repertoire with a primer and extending with a polymerase. This latter approach is described in Oliphant et al, Gene, 44: 177-183 (1986). Such duplex tags
- may then be inserted into cloning vectors along with target polynucleotides for sorting

20   and manipulation of the target polynucleotide in accordance with the invention.

When tag complements are employed that are made up of nucleotides that have enhanced binding characteristics, such as PNAs or oligonucleotide $N3' \rightarrow P5'$ phosphoramidates, sorting can be implemented through the formation of D-loops between tags comprising natural nucleotides and their PNA or phosphoramidate

25   complements, as an alternative to the "stripping" reaction employing the $3' \rightarrow 5'$ exonuclease activity of a DNA polymerase to render a tag single stranded.

Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length

30   from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers of subunits, these ranges may be expressed as follows:

Table IV
Numbers of Subunits in Tags in Preferred Embodiments

35

Monomers in Subunit	Nucleotides in Oligonucleotide Tag		
	(12-60)	(18-40)	(25-40)

3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

Most preferably, oligonucleotide tags are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

Preferably, repertoires of single stranded oligonucleotide tags of the invention

5   contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

Triplex Tags

In embodiments where specific hybridization occurs via triplex formation,

10   coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*"

15   indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate

20   sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Roberts et al, Proc. Natl.

25   Acad. Sci., 93: 4320-4325 (1996); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan,

30   Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem., 267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl.

32: 666-690 (1993); Escude et al, Proc. Natl. Acad. Sci., 93: 4365-4369 (1996); and
the like.   Conditions for annealing single-stranded or duplex tags to their single-
stranded or duplex complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-
1328 (1993); Cantor et al, U.S. patent 5,482,836; and the like.   Use of triplex tags has

5   the advantage of not requiring a "stripping" reaction with polymerase to expose the
tag for annealing to its complement.

Preferably, oligonucleotide tags of the invention employing triplex
hybridization are double stranded DNA and the corresponding tag complements are
single stranded.   More preferably, 5-methylcytosine is used in place of cytosine in the

10   tag complements in order to broaden the range of pH stability of the triplex formed
between a tag and its complement.   Preferred conditions for forming triplexes are
fully disclosed in the above references.   Briefly, hybridization takes place in
concentrated salt solution, e.g. 1.0 M NaCl, 1.0 M potassium acetate, or the like, at
pH below 5.5 (or 6.5 if 5-methylcytosine is employed).   Hybridization temperature

15   depends on the length and composition of the tag; however, for an 18-20-mer tag of
longer, hybridization at room temperature is adequate.   Washes may be conducted
with less concentrated salt solutions, e.g. 10 mM sodium acetate, 100 mM $MgCl_2$, pH
5.8, at room temperature.   Tags may be eluted from their tag complements by
- incubation in a similar salt solution at pH 9.0.

20   Minimally cross-hybridizing sets of oligonucleotide tags that form triplexes
may be generated by the computer program of Appendix Ic, or similar programs.   An
exemplary set of double stranded 8-mer words are listed below in capital letters with
the corresponding complements in small letters.   Each such word differs from each of
the other words in the set by three base pairs.

25

Table V

Exemplary Minimally Cross-Hybridizing
Set of DoubleStranded 8-mer Tags

5'-AAGGAGAG	5'-AAAGGGGA	5'-AGAGAAGA	5'-AGGGGGGG
3'-TTCCTCTC	3'-TTTCCCCT	3'-TCTCTTCT	3'-TCCCCCCC
3'-ttcctctc	3'-tttcccct	3'-tctcttct	3'-tccccccc
5'-AAAAAAAA	5'-AAGAGAGA	5'-AGGAAAAG	5'-GAAAGGAG
3'-TTTTTTTT	3'-TTCTCTCT	3'-TCCTTTTC	3'-CTTTCCTC
3'-tttttttt	3'-ttctctct	3'-tccttttc	3'-ctttcctc
5'-AAAAAGGG	5'-AGAAGAGG	5'-AGGAAGGA	5'-GAAGAAGG
3'-TTTTTCCC	3'-TCTTCTCC	3'-TCCTTCCT	3'-CTTCTTCC
3'-tttttccc	3'-tcttctcc	3'-tccttcct	3'-cttcttcc
5'-AAAGGAAG	5'-AGAAGGAA	5'-AGGGGAAA	5'-GAAGAGAA
3'-TTTCCTTC	3'-TCTTCCTT	3'-TCCCCTTT	3'-CTTCTCTT
3'-tttccttc	3'-tcttcctt	3'-tccccttt	3'-cttctctt

5

10

Table VI
Repertoire Size of Various Double Stranded Tags
That Form Triplexes with Their Tag Complements

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross- Hybridizing Set	Maximal Size of Minimally Cross- Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	2	8	4096	3.2×10^4
6	3	8	4096	3.2×10^4
8	3	16	6.5×10^4	1.05×10^6
10	5	8	4096	
15	5	92		
20	6	765		
20	8	92		
20	10	22		

15   Preferably, repertoires of double stranded oligonucleotide tags of the invention
contain at least 10 members; more preferably, repertoires of such tags contain at least
100 members. Preferably, words are between 4 and 8 nucleotides in length for
combinatorially synthesized double stranded oligonucletide tags, and oligonucleotide
tags are between 12 and 60 base pairs in length. More preferably, such tags are
20   between 18 and 40 base pairs in length.

Solid Phase Supports
        Solid phase supports for use with the invention may have a wide variety of
forms, including microparticles, beads, and membranes, slides, plates, micromachined
25   chips, and the like. Likewise, solid phase supports of the invention may comprise a

wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary

5   sequences of the same tag (and no other), or a single or a few supports are employed with spatially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several μm^2, e.g. 3-5, to several hundred μm^2, e.g. 100-

10   500. Preferably, such regions are spatially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

15   Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, Nucleic Acids Research, 16: 10861-10880 (1988); Albretsen et al, Anal. Biochem., 189: 40-50 (1990); Wolf et al, Nucleic Acids Research, 15: 2911-2926 (1987); or Ghosh et al, Nucleic Acids Research, 15:

20   5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including

25   microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: Meth. Enzymol., Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046,720; and Pon, Chapter 19, in Agrawal, editor, Methods in

30   Molecular Biology, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGel^{TM}, Rapp Polymere, Tubingen Germany); and the like. Selection of the

35   support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hindrance of the enzymes and that facilitate

access to substrate are preferred. Other important factors to be considered in selecting the most appropriate microparticle support include size uniformity, efficiency as a synthesis support, degree to which surface area known, and optical properties, e.g. as explain more fully below, clear smooth beads provide instrumentational advantages

5   when handling large numbers of beads on a surface.

Exemplary linking moieties for attaching and/or synthesizing tags on microparticle surfaces are disclosed in Pon et al, Biotechniques, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, J. Chem. Soc. Commun., 1989: 891-893; Damha et

10   al, Nucleic Acids Research, 18: 3813-3821 (1990); Beattie et al, Clinical Chemistry, 39: 719-722 (1993); Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag

15   complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern and Maskos, International application -PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, Genomics, 13:

20   1008-1017 (1992); and Maskos and Southern, Nucleic Acids Research, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces

25   are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, Oligonucleotide Synthesis: A Practical Approach (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000 μm diameter are preferable, as they facilitate the

30   construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

In some preferred applications, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached.

35   e.g. Applied Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large

numbers of microparticles on planar supports, such as a microscope slide. Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5 μm diameter GMA beads are employed.

Attaching Tags to Polynucleotides
For Sorting onto Solid Phase Supports

An important aspect of the invention is the sorting and attachment of a populations of polynucleotides, e.g. from a cDNA library, to microparticles or to separate regions on a solid phase support such that each microparticle or region has substantially only one kind of polynucleotide attached. This objective is accomplished by insuring that substantially all different polynucleotides have different tags attached. This condition, in turn, is brought about by taking a sample of the full ensemble of tag-polynucleotide conjugates for analysis. (It is acceptable that identical polynucleotides have different tags, as it merely results in the same polynucleotide being operated on or analyzed twice in two different locations.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the polynucleotides, it can be carried out inherently as a secondary effect of the techniques used to process the polynucleotides and tags, or sampling can be carried out both overtly and as an inherent part of processing steps.

Preferably, in constructing a cDNA library where substantially all different cDNAs have different tags, a tag repertoire is employed whose complexity, or number of distinct tags, greatly exceeds the total number of mRNAs extracted from a cell or tissue sample. Preferably, the complexity of the tag repertoire is at least 10 times that of the polynucleotide population; and more preferably, the complexity of the tag repertoire is at least 100 times that of the polynucleotide population. Below, a protocol is disclosed for cDNA library construction using a primer mixture that contains a full repertoire of exemplary 9-word tags. Such a mixture of tag-containing primers has a complexity of 8^9, or about 1.34×10^8. As indicated by Winslow et al, Nucleic Acids Research, 19: 3251-3253 (1991), mRNA for library construction can be extracted from as few as 10-100 mammalian cells. Since a single mammalian cell contains about 5×10^5 copies of mRNA molecules of about 3.4×10^4 different kinds,

by standard techniques one can isolate the mRNA from about 100 cells, or (theoretically) about 5×10^7 mRNA molecules. Comparing this number to the complexity of the primer mixture shows that without any additional steps, and even assuming that mRNAs are converted into cDNAs with perfect efficiency (1%

5   efficiency or less is more accurate), the cDNA library construction protocol results in a population containing no more than 37% of the total number of different tags. That is, without any overt sampling step at all, the protocol inherently generates a sample that comprises 37%, or less, of the tag repertoire. The probability of obtaining a double under these conditions is about 5%, which is within the preferred range. With

10 mRNA from 10 cells, the fraction of the tag repertoire sampled is reduced to only 3.7%, even assuming that all the processing steps take place at 100% efficiency. In fact, the efficiencies of the processing steps for constructing cDNA libraries are very low, a "rule of thumb" being that good library should contain about 10^8 cDNA clones from mRNA extracted from 10^6 mammalian cells.

15   Use of larger amounts of mRNA in the above protocol, or for larger amounts of polynucleotides in general, where the number of such molecules exceeds the complexity of the tag repertoire, a tag-polynucleotide conjugate mixture potentially contains every possible pairing of tags and types of mRNA or polynucleotide. In such cases, overt sampling may be implemented by removing a sample volume after a

20 serial dilution of the starting mixture of tag-polynucleotide conjugates. The amount of dilution required depends on the amount of starting material and the efficiencies of the processing steps, which are readily estimated.

        If mRNA were extracted from 10^6 cells (which would correspond to about 0.5 μg of poly(A)$^+$ RNA), and if primers were present in about 10-100 fold concentration

25 excess--as is called for in a typical protocol, e.g. Sambrook et al, Molecular Cloning, Second Edition, page 8.61 [10 μL 1.8 kb mRNA at 1 mg/mL equals about 1.68×10^{-11} moles and 10 μL 18-mer primer at 1 mg/mL equals about 1.68×10^{-9} moles], then the total number of tag-polynucleotide conjugates in a cDNA library would simply be equal to or less than the starting number of mRNAs, or about 5×10^{11} vectors

30 containing tag-polynucleotide conjugates--again this assumes that each step in cDNA construction--first strand synthesis, second strand synthesis, ligation into a vector-- occurs with perfect efficiency, which is a very conservative estimate. The actual number is significantly less.

        If a sample of n tag-polynucleotide conjugates are randomly drawn from a

35 reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution, $P(r)=e^{-\lambda}(\lambda)^r/r$, where r is the number of conjugates having the same tag and $\lambda=np$, where p is the probability of a given tag being selected. If $n=10^6$ and $p=1/(1.34 \times$

10^8), then $\lambda=.00746$ and $P(2)=2.76 \times 10^{-5}$. Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained as follows: Assume that the 5×10^{11} mRNAs are perfectly converted into 5×10^{11} vectors with tag-cDNA conjugates as inserts and that the $5 \times$

5   10^{11} vectors are in a reaction solution having a volume of 100 µl. Four 10-fold serial dilutions may be carried out by transferring 10 µl from the original solution into a vessel containing 90 µl of an appropriate buffer, such as TE. This process may be repeated for three additional dilutions to obtain a 100 µl solution containing 5×10^5 vector molecules per µl. A 2 µl aliquot from this solution yields 10^6 vectors

10   containing tag-cDNA conjugates as inserts. This sample is then amplified by straight forward transformation of a competent host cell followed by culturing.

Of course, as mentioned above, no step in the above process proceeds with perfect efficiency. In particular, when vectors are employed to amplify a sample of tag-polynucleotide conjugates, the step of transforming a host is very inefficient.

15   Usually, no more than 1% of the vectors are taken up by the host and replicated. Thus, for such a method of amplification, even fewer dilutions would be required to obtain a sample of 10^6 conjugates.

A repertoire of oligonucleotide tags can be conjugated to a population of polynucleotides in a number of ways, including direct enzymatic ligation,

20   amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, as noted above, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two

25   different polynucleotides, can be made negligible. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation or adequate representation of a rapidly changing mRNA pool,

30   and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored.

35   As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual

percentages of tag-molecule conjugates depends on how the tags are being employed. Preferably, for nucleic acid sequencing, substantially all means that at least eighty percent of the polynucleotides have unique tags attached. More preferably, it means that at least ninety percent of the polynucleotides have unique tags attached. Still

5   more preferably, it means that at least ninety-five percent of the polynucleotides have unique tags attached. And, most preferably, it means that at least ninety-nine percent of the polynucleotides have unique tags attached.

Preferably, when the population of polynucleotides consists of messenger RNA (mRNA), oligonucleotides tags may be attached by reverse transcribing the

10   mRNA with a set of primers preferably containing complements of tag sequences. An exemplary set of such primers could have the following sequence (SEQ ID NO: 1):

```
5'-mRNA- [A]n -3'
                [T]19GG[W,W,W,C]9ACCAGCTGATC-5'-biotin
```
15

where "[W,W,W,C]9" represents the sequence of an oligonucleotide tag of nine subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences

20   listed above, i.e. "W" represents T or A. The underlined sequences identify an optional restriction endonuclease site that can be used to release the polynucleotide from attachment to a solid phase support via the biotin, if one is employed. For the above primer, the complement attached to a microparticle could have the form:

25   5'-[G,W,W,W]9TGG-linker-microparticle

After reverse transcription, the mRNA is removed, e.g. by RNase H digestion. and the second strand of the cDNA is synthesized using, for example, a primer of the following form (SEQ ID NO: 2):

30

              5'-NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction

35   site in the resulting double stranded DNA which, together with the Sal I site, facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst Y1 and Sal I digestion, the exemplary conjugate would have the form:

```
5'-RCGACCA[C,W,W,W]9GG[T]19- cDNA -NNNR
      GGT[G,W,W,W]9CC[A]19- rDNA -NNNYCTAG-5'
```

The polynucleotide-tag conjugates may then be manipulated using standard molecular

5 biology techniques. For example, the above conjugate--which is actually a mixture--
may be inserted into commercially available cloning vectors, e.g. Stratagene Cloning
System (La Jolla, CA); transfected into a host, such as a commercially available host
bacteria; which is then cultured to increase the number of conjugates. The cloning
vectors may then be isolated using standard techniques, e.g. Sambrook et al,

10 Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York,
1989). Alternatively, appropriate adaptors and primers may be employed so that the
conjugate population can be increased by PCR.

Preferably, when the ligase-based method of sequencing is employed, the Bst
Y1 and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector

15 having the following single-copy restriction sites (SEQ ID NO: 3):


```
5'-GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA-3'
    FokI          BamHI  XhoI
```

20

This adds the Fok I site which will allow initiation of the sequencing process
discussed more fully below.

Tags can be conjugated to cDNAs of existing libraries by standard cloning
methods. cDNAs are excised from their existing vector, isolated, and then ligated into

25 a vector containing a repertoire of tags. Preferably, the tag-containing vector is
linearized by cleaving with two restriction enzymes so that the excised cDNAs can be
ligated in a predetermined orientation. The concentration of the linearized tag-
containing vector is in substantial excess over that of the cDNA inserts so that
ligation provides an inherent sampling of tags.

30   A general method for exposing the single stranded tag after amplification
involves digesting a target polynucleotide-containing conjugate with the 5'→3'
exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the
presence of a single deoxynucleoside triphosphate, such a polymerase will cleave
nucleotides from 3' recessed ends present on the non-template strand of a double

35 stranded fragment until a complement of the single deoxynucleoside triphosphate is
reached on the template strand. When such a nucleotide is reached the 5'→3'
digestion effectively ceases, as the polymerase's extension activity adds nucleotides at
a higher rate than the excision activity removes nucleotides. Consequently, single

stranded tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the
5   polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site. Clearly, this procedure can be generalized to
10   endonucleases other than Fok I.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags
15   and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Sambrook et al, Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions
20   are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

25   When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the
30   relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing
35   on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range

of 20-50 μm are loaded with about 10^5 polynucleotides, and GMA beads of diameter in the range of 5-10 μm are loaded with a few tens of thousand of polynucleotides, e.g. 4×10^4 to 6×10^4.

5   In the preferred embodiment, tag complements are synthesized on microparticles combinatorially; thus, at the end of the synthesis, one obtains a complex mixture of microparticles from which a sample is taken for loading tagged polynucleotides. The size of the sample of microparticles will depend on several factors, including the size of the repertoire of tag complements, the nature of the apparatus for used for observing loaded microparticles--e.g. its capacity, the tolerance 10 for multiple copies of microparticles with the same tag complement (i.e. "bead doubles"), and the like. The following table provide guidance regarding microparticle sample size, microparticle diameter, and the approximate physical dimensions of a packed array of microparticles of various diameters.

15

Microparticle diameter	5 μm	10 μm	20 μm	40 μm
Max. no. polynucleotides loaded at 1 per 10^5 sq. angstrom		3×10^5	1.26×10^6	5×10^6
Approx. area of monolayer of 10^6 microparticles	.45 x .45 cm	1 x 1 cm	2 x 2 cm	4 x 4 cm

20   The probability that the sample of microparticles contains a given tag complement or is present in multiple copies is described by the Poisson distribution, as indicated in the following table.

25

Table VII

Number of microparticles in sample (as fraction of repertoire size), m	Fraction of repertoire of tag complements present in sample, $1-e^{-m}$	Fraction of microparticles in sample with unique tag complement attached, $m(e^{-m})/2$	Fraction of microparticles in sample carrying same tag complement as one other microparticle in sample ("bead doubles"), $m^2(e^{-m})/2$
1.000	0.63	0.37	0.18
.693	0.50	0.35	0.12
.405	0.33	0.27	0.05
.285	0.25	0.21	0.03
.223	0.20	0.18	0.02
.105	0.10	0.09	0.005
.010	0.01	0.01	

High Specificity Sorting and Panning

The kinetics of sorting depends on the rate of hybridization of oligonucleotide tags to their tag complements which, in turn, depends on the complexity of the tags in the hybridization reaction. Thus, a trade off exists between sorting rate and tag complexity, such that an increase in sorting rate may be achieved at the cost of reducing the complexity of the tags involved in the hybridization reaction. As explained below, the effects of this trade off may be ameliorated by "panning."

Specificity of the hybridizations may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words selected from Table II, a repertoire of 8^8, or about 1.67×10^7, tags and tag complements are produced. In a library of tag-cDNA conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly onto a repertoire-equivalent of microparticles, or in this example a sample of 1.67×10^7 microparticles, then only a sparse subset of the sampled microparticles would be loaded. The density of loaded microparticles can be increase--for example, for more efficient sequencing--by undertaking a "panning" step in which the sampled tag-cDNA conjugates are used to separate loaded microparticles from unloaded microparticles. Thus, in the example above, even though a "0.1 percent" sample

contains only 16,700 cDNAs, the sampling and panning steps may be repeated until as many loaded microparticles as desired are accumulated.

A panning step may be implemented by providing a sample of tag-cDNA conjugates each of which contains a capture moiety at an end opposite, or distal to,
5   the oligonucleotide tag. Preferably, the capture moiety is of a type which can be released from the tag-cDNA conjugates, so that the tag-cDNA conjugates can be sequenced with a single-base sequencing method. Such moieties may comprise biotin, digoxigenin, or like ligands, a triplex binding region, or the like. Preferably, such a capture moiety comprises a biotin component. Biotin may be attached to tag-
10 cDNA conjugates by a number of standard techniques. If appropriate adapters containing PCR primer binding sites are attached to tag-cDNA conjugates, biotin may be attached by using a biotinylated primer in an amplification after sampling. Alternatively, if the tag-cDNA conjugates are inserts of cloning vectors, biotin may be attached after excising the tag-cDNA conjugates by digestion with an appropriate
15 restriction enzyme followed by isolation and filling in a protruding strand distal to the tags with a DNA polymerase in the presence of biotinylated uridine triphosphate.

After a tag-cDNA conjugate is captured, it may be released from the biotin moiety in a number of ways, such as by a chemical linkage that is cleaved by
-reduction, e.g. Herman et al, Anál. Biochem., 156: 48-55 (1986), or that is cleaved
20 photochemically, e.g. Olejnik et al, Nucleic Acids Research, 24: 361-366 (1996), or that is cleaved enzymatically by introducing a restriction site in the PCR primer. The latter embodiment can be exemplified by considering the library of tag-polynucleotide conjugates described above:

25   5'-RCGACCA[C,W,W,W]9GG[T]19- cDNA -NNNR
                 GGT[G,W,W,W]9CC[A]19- rDNA -NNNYCTAG-5'

The following adapters may be ligated to the ends of these fragments to permit amplification by PCR:
30

         5'- XXXXXXXXXXXXXXXXXXXX
             XXXXXXXXXXXXXXXXXXXXYGAT

35   Right Adapter

         GATCZZACTAGTZZZZZZZZZZZZ-3'
40   ZZTGATCAZZZZZZZZZZZZ

Left Adapter

```
ZZTGATCAZZZZZZZZZZZZ-5'-biotin
```

5

Left Primer

where "ACTAGT" is a Spe I recognition site (which leaves a staggered cleavage
ready for single base sequencing), and the X's and Z's are nucleotides selected so that
10   the annealing and dissociation temperatures of the respective primers are
approximately the same. After ligation of the adapters and amplification by PCR
using the biotinylated primer, the tags of the conjugates are rendered single stranded
by the exonuclease activity of T4 DNA polymerase and conjugates are combined with
a sample of microparticles, e.g. a repertoire equivalent, with tag complements
15   attached. After annealing under stringent conditions (to minimize mis-attachment of
tags), the conjugates are preferably ligated to their tag complements and the loaded
microparticles are separated from the unloaded microparticles by capture with
avidinated magnetic beads, or like capture technique.

Returning to the example, this process results in the accumulation of about
20   10,500 (=16,700 x .63) loaded microparticles with different tags, which may be
released from the magnetic beads by cleavage with Spe I. By repeating this process
40-50 times with new samples of microparticles and tag-cDNA conjugates, 4-5 x 10^5
cDNAs can be accumulated by pooling the released microparticles. The pooled
microparticles may then be simultaneously sequenced by a single-base sequencing
25   technique.

Determining how many times to repeat the sampling and panning steps--or
more generally, determining how many cDNAs to analyze, depends on one's
objective. If the objective is to monitor the changes in abundance of relatively
common sequences, e.g. making up 5% or more of a population, then relatively small
30   samples, i.e. a small fraction of the total population size, may allow statistically
significant estimates of relative abundances. On the other hand, if one seeks to
monitor the abundances of rare sequences, e.g. making up 0.1% or less of a
population, then large samples are required. Generally, there is a direct relationship
between sample size and the reliability of the estimates of relative abundances based
35   on the sample. There is extensive guidance in the literature on determining
appropriate sample sizes for making reliable statistical estimates, e.g. Koller et al,
Nucleic Acids Research, 23:185-191 (1994); Good, Biometrika, 40: 16-264 (1953);
Bunge et al, J. Am. Stat. Assoc., 88: 364-373 (1993); and the like. Preferably, for

monitoring changes in gene expression based on the analysis of a series of cDNA libraries containing 10^5 to 10^8 independent clones of 3.0-3.5×10^4 different sequences, a sample of at least 10^4 sequences are accumulated for analysis of each library. More preferably, a sample of at least 10^5 sequences are accumulated for the

5   analysis of each library; and most preferably, a sample of at least 5×10^5 sequences are accumulated for the analysis of each library. Alternatively, the number of sequences sampled is preferably sufficient to estimate the relative abundance of a sequence present at a frequency within the range of 0.1% to 5% with a 95% confidence limit no larger than 0.1% of the population size.

10

Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple

15   polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably, the methodology permits the stepwise identification of nucleotides, usually one at a - time, in a sequence through successive cycles of treatment and detection. Such

20   methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al, International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

25   A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. Briefly, the method comprises the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a

30   complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide,

35   or a portion thereof, is determined.

A single signal generating moiety, such as a single fluorescent dye, may be employed when sequencing several different target polynucleotides attached to different spatially addressable solid phase supports, such as fixed microparticles, in a

parallel sequencing operation. This may be accomplished by providing four sets of probes that are applied sequentially to the plurality of target polynucleotides on the different microparticles. An exemplary set of such probes are shown below:

5

Set 1	Set 2	Set 3	Set 4
ANNNN...NN	dANNNN...NN	dANNNN...NN	dANNNN...NN
N...NNTT...T*	d N...NNTT...T	N...NNTT...T	N...NNTT...T
dCNNNN...NN	CNNNN...NN	dCNNNN...NN	dCNNNN...NN
N...NNTT...T	N...NNTT...T*	N...NNTT...T	N...NNTT...T
dGNNNN...NN	dGNNNN...NN	GNNNN...NN	dGNNNN...NN
N...NNTT...T	N...NNTT...T	N...NNTT...T*	N...NNTT...T
dTNNNN...NN	dTNNNN...NN	dTNNNN...NN	TNNNN...NN
N...NNTT...T	N...NNTT...T	N...NNTT...T	N...NNTT...T*

where each of the listed probes represents a mixture of $4^3=64$ oligonucleotides such that the identity of the 3' terminal nucleotide of the top strand is fixed and the other positions in the protruding strand are filled by every 3-mer permutation of nucleotides,

10   or complexity reducing analogs. The listed probes are also shown with a single stranded poly-T tail with a signal generating moiety attached to the terminal thymidine, shown as "T*". The "d" on the unlabeled probes designates a ligation-blocking moiety or absense of 3'-hydroxyl, which prevents unlabeled probes from being ligated. Preferably, such 3'-terminal nucleotides are dideoxynucleotides. In this embodiment,

15   the probes of set 1 are first applied to the plurality of target polynucleotides and treated with a ligase so that target polynucleotides having a thymidine complementary to the 3' terminal adenosine of the labeled probes are ligated. The unlabeled probes are simultaneously applied to minimize inappropriate ligations. The locations of the target polynucleotides that form ligated complexes with probes terminating in "A" are

20   identified by the signal generated by the label carried on the probe. After washing and cleavage, the probes of set 2 are applied. In this case, target polynucleotides forming ligated complexes with probes terminating in "C" are identified by location. Similarly, the probes of sets 3 and 4 are applied and locations of positive signals identified. This process of sequentially applying the four sets of probes continues until the desired

25   number of nucleotides are identified on the target polynucleotides. Clearly, one of ordinary skill could construct similar sets of probes that could have many variations, such as having protruding strands of different lengths, different moieties to block ligation of unlabeled probes, different means for labeling probes, and the like.

Apparatus for Sequencing Populations of Polynucleotides

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the

5   molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide

10 sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

15 Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent applications - PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system

20 should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available

25 components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is

30 illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available form Apple Computer (Cupertino, CA).

35 Computer software for table translation and data collection functions can be provided by commercially available laboratory software, such as Lab Windows, available from National Instruments.

Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a

5   dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or more suitable opto-electronic devices for converting some fluorescence characteristic,

10   e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 2, signal distribution optics 318 directs fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336,

15   whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD
  - array.

20   The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least

25   have the capability of resolving objects on the order of 10-100 μm. Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where

30   microparticles are known to be located, e.g from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per cm^2.

In sequencing applications, loaded microparticles can be fixed to the surface

35   of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which

in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique

5 restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the

10 associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the

15 initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type IIs restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture comprises about 10-15 percent of the biotinylated probe.

20 In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles. Preferably, such a glass substrate is a flow cell, which may comprise a channel etched

25 in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

Identification of Novel Polynucleotides
in cDNA Libraries

30 Novel polynucleotides in a cDNA library can be identified by constructing a library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as

35 taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like references, the following primer may by hybridized to the polyA tails for first strand synthesis with a reverse transcriptase using conventional protocols (SEQ ID NO: 1):

```
5'-mRNA- [A]n -3'
          [T]19-[primer site]-GG[W,W,W,C]9ACCAGCTGATC-5'
```

where $[W,W,W,C]_9$ represents a tag as described above, "ACCAGCTGATC" is an
5   optional sequence forming a restriction site in double stranded form, and "primer site"
is a sequence common to all members of the library that is later used as a primer
binding site for amplifying polynucleotides of interest by PCR.

After reverse transcription and second strand synthesis by conventional
techniques, the double stranded fragments are inserted into a cloning vector as
10   described above and amplified. The amplified library is then sampled and the sample
amplified. The cloning vectors from the amplified sample are isolated, and the tagged
cDNA fragments excised and purified. After rendering the tag single stranded with a
polymerase as described above, the fragments are methylated and sorted onto
microparticles in accordance with the invention. Preferably, as described above, the
15   cloning vector is constructed so that the tagged cDNAs can be excised with an
endonuclease, such as Fok I, that will allow immediate sequencing by the preferred
single base method after sorting and ligation to microparticles.

Stepwise sequencing is then carried out simultaneously on the whole library,
or one or more large fractions of the library, in accordance with the invention until a
20   sufficient number of nucleotides are identified on each cDNA for unique
representation in the genome of the organism from which the library is derived. For
example, if the library is derived from mammalian mRNA then a randomly selected
sequence 14-15 nucleotides long is expected to have unique representation among the
2-3 thousand megabases of the typical mammalian genome. Of course identification
25   of far fewer nucleotides would be sufficient for unique representation in a library
derived from bacteria, or other lower organisms. Preferably, at least 20-30
nucleotides are identified to ensure unique representation and to permit construction
of a suitable primer as described below. The tabulated sequences may then be
compared to known sequences to identify unique cDNAs.
30   Unique cDNAs are then isolated by conventional techniques, e.g. constructing
a probe from the PCR amplicon produced with primers directed to the prime site and
the portion of the cDNA whose sequence was determined. The probe may then be
used to identify the cDNA in a library using a conventional screening protocol.

The above method for identifying new cDNAs may also be used to fingerprint
35   mRNA populations, either in isolated measurements or in the context of a
dynamically changing population. Partial sequence information is obtained
simultaneously from a large sample, e.g. ten to a hundred thousand, or more, of
cDNAs attached to separate microparticles as described in the above method.

Example 1
Construction of a Tag Library

An exemplary tag library is constructed as follows to form the chemically
synthesized 9-word tags of nucleotides A, G, and T defined by the formula:

$$3'\text{-TGGC-}[^4(A,G,T)_9]\text{-CCCCp}$$

where "$[^4(A,G,T)_9]$" indicates a tag mixture where each tag consists of nine 4-mer
words of A, G, and T; and "p" indicate a 5' phosphate. This mixture is ligated to the
following right and left primer binding regions (SEQ ID NO: 4 and SEQ ID NO 5):

```
5'- AGTGGCTGGGCATCGGACCG          5'- GGGGCCCAGTCAGCGTCGAT
    TCACCGACCCGTAGCCp                  GGGTCAGTCGCAGCTA
```

                    LEFT   RIGHT

The right and left primer binding regions are ligated to the above tag mixture, after
which the single stranded portion of the ligated structure is filled with DNA
polymerase then mixed with the right and left primers indicated below and amplified
to give a tag library (SEQ ID NO: 6).

```
        Left Primer

5'- AGTGGCTGGGCATCGGACCG

5'- AGTGGCTGGGCATCGGACCG-  [4(A,G,T)9]-GGGGCCCAGTCAGCGTCGAT
    TCACCGACCCGTAGCCTGGC-  [4(A,G,T)9]-CCCCGGGTCAGTCGCAGCTA

                               CCCCGGGTCAGTCGCAGCTA-5'

                                   Right Primer
```

The underlined portion of the left primer binding region indicates a Rsr II recognition
site. The left-most underlined region of the right primer binding region indicates
recognition sites for Bsp 120I, Apa I, and Eco O 109I, and a cleavage site for Hga I.
The right-most underlined region of the right primer binding region indicates the
recognition site for Hga I. Optionally, the right or left primers may be synthesized
with a biotin attached (using conventional reagents, e.g. available from Clontech
Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or
cleavage.

NOT FURNISHED UPON FILING

```
                            primer binding site              Ppu MI site
                                   ↓                             ↓
              (plasmid)-5' -AAAAGGAGGAGGCCTTGATAGAGAGGACCT-
                           -TTTTCCTCCTCCGGAACTATCTCTCCTGGA-

                                  primer binding site
                                          ↓
             -GTTTAAAC-GGATCC-TCTTCCTCTTCCTCTTCC-3'-(plasmid)
             -CAAATTTG-CCTAGG-AGAAGGAGAAGGAGAAGG-
                  ↑        ↑
                       Bam HI site
          Pme I site
```

The plasmid is cleaved with Ppu MI and Pme I (to give a Rsr II-compatible end and a flush end so that the insert is oriented) and then methylated with DAM methylase. The tag-containing construct is cleaved with Rsr II and then ligated to the open plasmid, after which the conjugate is cleaved with Mbo I and Bam HI to permit ligation and closing of the plasmid. The plasmid is then amplified and isolated and used in accordance with the invention.

Example 3
Changes in Gene Expression Profiles in Liver Tissue of Rats
Exposed to Various Xenobiotic Agents

In this experiment, to test the capability of the method of the invention to detect genes induced as a result of exposure to xenobiotic compounds, the gene expression profile of rat liver tissue is examined following administration of several compounds known to induce the expression of cytochrome P-450 isoenzymes. The results obtained from the method of the invention are compared to results obtained from reverse transcriptase PCR measurements and immunochemical measurements of the cytochrome P-450 isoenzymes. Protocols and materials for the latter assays are described in Morris et al, Biochemical Pharmacology, 52: 781-792 (1996).

Male Sprague-Dawley rats between the ages of 6 and 8 weeks and weighing 200-300 g are used, and food and water are available to the animals *ad lib*. Test compounds are phenobarbital (PB), metyrapone (MET), dexamethasone (DEX), clofibrate (CLO), corn oil (CO), and β-naphthoflavone (BNF), and are available from Sigma Chemical Co. (St. Louis, MO). Antibodies against specific P-450 enzymes are available from the following sources: rabbit anti-rat CYP3A1 from Human Biologics, Inc. (Phoenix, AZ); goat anti-rat CYP4A1 from Daiichi Pure Chemicals Co. (Tokyo,

Japan); monoclonal mouse anti-rat CYP1A1, monoclonal mouse anti-rat CYP2C11, goat anti-rat CYP2E1, and monoclonal mouse anti-rat CYP2B1 from Oxford Biochemical Research, Inc. (Oxford, MI). Secondary antibodies (goat anti-rabbit IgG, rabbit anti-goat IgG and goat anti-mouse IgG) are available from Jackson

5   ImmunoResearch Laboratories (West Grove, PA).

Animals are administered either PB (100 mg/kg), BNF (100 mg/kg), MET (100 mg/kg), DEX (100 mg/kg), or CLO (250 mg/kg) for 4 consecutive days via intraperitoneal injection following a dosing regimen similar to that described by Wang et al, Arch. Biochem. Biophys. 290: 355-361 (1991). Animals treated with

10   H_2O and CO are used as controls. Two hours following the last injection (day 4), animals are killed, and the livers are removed. Livers are immediately frozen and stored at -70°C.

Total RNA is prepared from frozen liver tissue using a modification of the method described by Xie et al, Biotechniques, 11: 326-327 (1991). Approximately

15   100-200 mg of liver tissue is homogenized in the RNA extraction buffer described by Xie et al to isolate total RNA. The resulting RNA is reconstituted in diethylpyrocarbonate-treated water, quantified spectrophotometrically at 260 nm, and adjusted to a concentration of 100 µg/ml. Total RNA is stored in
- diethylpyrocarbonate-treated water for up to 1 year at -70°C without any apparent

20   degradation. RT-PCR and sequencing are performed on samples from these preparations.

For sequencing, samples of RNA corresponding to about 0.5 µg of poly(A)$^+$ RNA are used to construct libraries of tag-cDNA conjugates following the protocol described in the section entitled "Attaching Tags to Polynucleotides for Sorting onto

25   Solid Phase Supports," with the following exception: the tag repertoire is constructed from six 4-nucleotide words from Table II. Thus, the complexity of the repertoire is 8^6 or about 2.6×10^5. For each tag-cDNA conjugate library constructed, ten samples of about ten thousand clones are taken for amplification and sorting. Each of the amplified samples is separately applied to a fixed monolayer of about 10^6 10 µm

30   diameter GMA beads containing tag complements. That is, the "sample" of tag complements in the GMA bead population on each monolayer is about four fold the total size of the repertoire, thus ensuring there is a high probability that each of the sampled tag-cDNA conjugates will find its tag complement on the monolayer. After the oligonucleotide tags of the amplified samples are rendered single stranded as

35   described above, the tag-cDNA conjugates of the samples are separately applied to the monolayers under conditions that permit specific hybridization only between oligonucleotide tags and tag complements forming perfectly matched duplexes. Concentrations of the amplified samples and hybridization times are selected to

permit the loading of about 5×10^4 to 2×10^5 tag-cDNA conjugates on each bead where perfect matches occur. After ligation, 9-12 nucleotide portions of the attached cDNAs are determined in parallel by the single base sequencing technique described by Brenner in International patent application PCT/US95/03678. Frequency

5   distributions for the gene expression profiles are assembled from the sequence information obtained from each of the ten samples.

RT-PCRs of selected mRNAs corresponding to cytochrome P-450 genes and the constitutively expressed cyclophilin gene are carried out as described in Morris et al (cited above). Briefly, a 20 μL reaction mixture is prepared containing 1x reverse

10   transcriptase buffer (Gibco BRL), 10 nM dithiothreitol, 0.5 nM dNTPs, 2.5 μM oligo $d(T)_{15}$ primer, 40 units RNasin (Promega, Madison, WI), 200 units RNase H-reverse transcriptase (Gibco BRL), and 400 ng of total RNA (in diethylpyrocarbonate-treated water). The reaction is incubated for 1 hour at 37°C followed by inactivation of the enzyme at 95°C for 5 min. The resulting cDNA is stored at -20°C until used. For

15   PCR amplification of cDNA, a 10 μL reaction mixture is prepared containing 10x polymerase reaction buffer, 2 mM $MgCl_2$, 1 unit Taq DNA polymerase (Perkin-Elmer, Norwalk, CT), 20 ng cDNA, and 200 nM concentration of the 5' and 3' specific PCR primers of the sequences described in Morris et al (cited above). PCRs -are carried out in a Perkin-Elmer 9600 thermal cycler for 23 cycles using melting,

20   annealing, and extension conditions of 94°C for 30 sec., 56°C for 1 min., and 72°C for 1 min.. respectively. Amplified cDNA products are separated by PAGE using 5% native gels. Bands are detected by staining with ethidium bromide.

Western blots of the liver proteins are carried out using standard protocols after separation by SDS-PAGE. Briefly, proteins are separated on 10% SDS-PAGE

25   gels under reducing conditions and immunoblotted for detection of P-450 isoenzymes using a modification of the methods described in Harris et al, Proc. Natl. Acad. Sci., 88: 1407-1410 (1991). Protein are loaded at 50 μg/lane and resolved under constant current (250 V) for approximately 4 hours at 2°C. Proteins are transferred to nitrocellulose membranes (Bio-Rad, Hercules, CA) in 15 mM Tris buffer containing

30   120 mM glycine and 20% (v/v) methanol. The nitrocellulose membranes are blocked with 2.5% BSA and immunoblotted for P-450 isoenzymes using primary monoclonal and polyclonal antibodies and secondary alkaline phosphatase conjugated anti-IgG. Immunoblots are developed with the Bio-Rad alkaline phosphatase substrate kit.

The three types of measurements of P-450 isoenzyme induction showed

35   substantial agreement.

APPENDIX Ia

Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```
Program minxh
c
c
c
        integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
        dimension nbase(6)
c
c
        write(*,*)'ENTER SUBUNIT LENGTH'
        read(*,100)nsub
100     format(i1)
        open(1,file='sub4.dat',form='formatted',status='new')
c
c
      nset=0
      do 7000 m1=1,3
         do 7000 m2=1,3
            do 7000 m3=1,3
               do 7000 m4=1,3
                  sub1(1)=m1
                  sub1(2)=m2
                  sub1(3)=m3
                  sub1(4)=m4
c
c
        ndiff=3
c
c
c                  Generate set of subunits differing from
c                  sub1 by at least ndiff nucleotides.
c                  Save in mset1.
c
c
        jj=1
        do 900 j=1,nsub
900        mset1(1,j)=sub1(j)
c
c
        do 1000 k1=1,3
           do 1000 k2=1,3
              do 1000 k3=1,3
                 do 1000 k4=1,3
c
c
                    nbase(1)=k1
                    nbase(2)=k2
                    nbase(3)=k3
                    nbase(4)=k4
```

```
c
                n=0
                do 1200 j=1,nsub
                    if(sub1(j).eq.1 .and. nbase(j).ne.1 .or.
        1               sub1(j).eq.2 .and. nbase(j).ne.2 .or.
        3               sub1(j).eq.3 .and. nbase(j).ne.3) then
                    n=n+1
                    endif
1200            continue
c
c
        if(n.ge.ndiff) then
c
c
c                                           If number of mismatches
c                                           is greater than or equal
c                                           to ndiff then record
c                                           subunit in matrix mset
c
c
                jj=jj+1
                 do 1100 i=1,nsub
1100                mset1(jj,i)=nbase(i)
                endif
c
c
1000    continue
c
c
                do 1325 j2=1,nsub
                mset2(1,j2)=mset1(1,j2)
1325            mset2(2,j2)=mset1(2,j2)
c
c
c                                           Compare subunit 2 from
c                                           mset1 with each successive
c                                           subunit in mset1, i.e. 3,
c                                           4,5, ... etc. Save those
c                                           with mismatches .ge. ndiff
c                                           in matrix mset2 starting at
c                                           position 2.
c                                             Next transfer contents
c                                           of mset2 into mset1 and
c                                           start
c                                           comparisons again this time
c                                           starting with subunit 3.
c                                           Continue until all subunits
c                                           undergo the comparisons.
c
c
                npass=0
c
c
1700    continue
        kk=npass+2
        npass=npass+1
c
```

```
c
                do 1500 m=npass+2,jj
                   n=0
                   do 1600 j=1,nsub
                      if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
        2                mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
        2                mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
                         n=n+1
                      endif
1600               continue
                   if(n.ge.ndiff) then
                        kk=kk+1
                        do 1625 i=1,nsub
1625                        mset2(kk,i)=mset1(m,i)
                   endif
1500            continue
c
c                                                kk is the number of subunits
c                                                stored in mset2
c
c                                                Transfer contents of mset2
c                                                into mset1 for next pass.
c
c

                do 2000 k=1,kk
                   do 2000 m=1,nsub
2000                  mset1(k,m)=mset2(k,m)
                if(kk.lt.jj) then
                   jj=kk
                   goto 1700
                   endif
c
c
                 nset=nset+1
                write(1,7009)
7009             format(/)
                do 7008 k=1,kk
7008               write(1,7010)(mset1(k,m),m=1,nsub)
7010            format(4i1)
                write(*,*)
                write(*,120) kk,nset
120             format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
7000             continue
                 close(1)
c
c
                end
c               ****************************************
c               ****************************************
```

APPENDIX Ib

Exemplary computer program for generating

minimally cross hybridizing sets

(single stranded tag/single stranded tag complement)

```
Program tagN
c
c
c         Program tagN generates minimally cross-hybridizing
c           sets of subunits given i) N--subunit length, and ii)
c           an initial subunit sequence.  tagN assumes that only
c           3 of the four natural nucleotides are used in the tags.
c
c

          character*1 sub1(20)
          integer*2 mset(10000,20), nbase(20)
c
c

          write(*,*)'ENTER SUBUNIT LENGTH'
          read(*,100)nsub
100       format(i2)
c
c

          write(*,*)'ENTER SUBUNIT SEQUENCE'
          read(*,110)(sub1(k),k=1,nsub)
110       format(20a1)
c
c

          ndiff=10
c
c
c                 Let a=1 c=2 g=3 & t=4
c
c

          do 800 kk=1,nsub
          if(sub1(kk).eq.'a') then
             mset(1,kk)=1
             endif
                  if(sub1(kk).eq.'c') then
                     mset(1,kk)=2
                     endif
                        if(sub1(kk).eq.'g') then
                           mset(1,kk)=3
                           endif
                              if(sub1(kk).eq.'t') then
                                 mset(1,kk)=4
                                 endif
800       continue
c
c
c                 Generate set of subunits differing from .
c                 sub1 by at least ndiff nucleotides.
c
c
c

          jj=1
c
c

          do 1000 k1=1,3
```

```
                     do 1000 k2=1,3
                       do 1000 k3=1,3
                         do 1000 k4=1,3
                           do 1000 k5=1,3
                             do 1000 k6=1,3
                               do 1000 k7=1,3
                                 do 1000 k8=1,3
                                   do 1000 k9=1,3
                                     do 1000 k10=1,3
                 do 1000 k11=1,3
                   do 1000 k12=1,3
                     do 1000 k13=1,3
                       do 1000 k14=1,3
                         do 1000 k15=1,3
                           do 1000 k16=1,3
                             dc 1000 k17=1,3
                               do 1000 k18=1,3
                                 do 1000 k19=1,3
                                   do 1000 k20=1,3
c
c

                 nbase(1)=k1
                 nbase(2)=k2
                 nbase(3)=k3
                 nbase(4)=k4
                 nbase(5)=k5
                 nbase(6)=k6
                 nbase(7)=k7
                 nbase(8)=k8
                 nbase(9)=k9
                 nbase(10)=k10
                 nbase(11)=k11
                 nbase(12)=k12
                 nbase(13)=k13
                 nbase(14)=k14
                 nbase(15)=k15
                 nbase(16)=k16
                 nbase(17)=k17
                 nbase(18)=k18
                 nbase(19)=k19
                 nbase(20)=k20
c
c

             do 1250 nn=1,jj
c
              n=0
              do 1200 j=1,nsub
                if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1             mset(nn,j).eq.2 .and. nbase(j).ne.2 ..or.
     2             mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3             mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                  n=n+1
                endif
1200          continue
c
c
          if(n.lt.ndiff) then
             goto 1000
             endif
1250      continue
c
c
              jj=jj+1
              write(*,130)(nbase(i),i=1,nsub),jj
               do 1100 i=1,nsub
```

```
                            mset(jj,i)=nbase(i)
      1100              continue
      c
      c
      1000      continue
      c
      c
                write(*,*)
      130       format(10x,20(1x,i1),5x,i5)
                write(*,*)
                write(*,120) jj
      120       format(1x,'Number of words=',i5)
      c
      c
                end
      c
      c         ********************************************
      c         ********************************************
```

APPENDIX lc

Exemplary computer program for generating

minimally cross hybridizing sets

(double stranded tag/single stranded tag complement)

```
Program 3tagN
c
c
c          Program 3tagN generates minimally cross-hybridizing
c             sets of duplex subunits given i) N--subunit length,
c             and ii) an initial homopurine sequence.
c
c
          character*1 subl(20)
          integer*2 mset(10000,20), nbase(20)
c
c
          write(*,*)'ENTER SUBUNIT LENGTH'
          read(*,100)nsub
100       format(i2)
c
c
          write(*,*)'ENTER SUBUNIT SEQUENCE a & g only'
          read(*,110)(subl(k),k=1,nsub)
110       format(20al)
c
          ndiff=10
c
c                    Let a=1 and g=2
c
          do 800 kk=1,nsub
          if(subl(kk).eq.'a') then
             mset(1,kk)=1
           endif
              if(subl(kk).eq.'g') then
                 mset(1,kk)=2
                 endif
800       continue
c


          jj=1
c
          do 1000 k1=1,3
            do 1000 k2=1,3
              do 1000 k3=1,3
                do 1000 k4=1,3
                  do 1000 k5=1,3
                    do 1000 k6=1,3
                      do 1000 k7=1,3
                        do 1000 k8=1,3
                          do 1000 k9=1,3
                            do 1000 k10=1,3
          do 1000 k11=1,3
            do 1000 k12=1,3
              do 1000 k13=1,3
                do 1000 k14=1,3
                  do 1000 k15=1,3
                    do 1000 k16=1,3
                      do 1000 k17=1,3
                        do 1000 k18=1,3
```

```
                                do 1000 k19=1,3
                                 do 1000 k20=1,3
c
                        nbase(1)=k1
                        nbase(2)=k2
                        nbase(3)=k3
                        nbase(4)=k4
                        nbase(5)=k5
                        nbase(6)=k6
                        nbase(7)=k7
                        nbase(8)=k8
                        nbase(9)=k9
                        nbase(10)=k10
                        nbase(11)=k11
                        nbase(12)=k12
                        nbase(13)=k13
                        nbase(14)=k14
                        nbase(15)=k15
                        nbase(16)=k16
                        nbase(17)=k17
                        nbase(18)=k18
                        nbase(19)=k19
                        nbase(20)=k20
c
                    do 1250 nn=1,jj
c
                     n=0
                     do 1200 j=1,nsub
                        if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                     mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
     2                     mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3                     mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                           n=n+1
                        endif
1200                    continue
c
                if(n.lt.ndiff) then
                    goto 1000
                    endif
1250            continue
c
                    jj=jj+1
                    write(*,130)(nbase(i),i=1,nsub),jj
                     do 1100 i=1,nsub
                        mset(jj,i)=nbase(i)
1100                        continue
c
1000            continue
c
                write(*,*)
130             format(10x,20(1x,i1),5x,i5)
                write(*,*)
                write(*,120) jj
120             format(1x,'Number of words=',i5)
c
c
                end
```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: David W. Martin, Jr.

(ii) TITLE OF INVENTION: Measurement of Gene Expression profiles in
Toxicity Determination

(iii) NUMBER OF SEQUENCES: 7

(iv) CORRESPONDENCE ADDRESS:
        (A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics, Inc.
        (B) STREET: 3832 Bay Center Place
        (C) CITY: Hayward
        (D) STATE: California
        (E) COUNTRY: USA
        (F) ZIP: 94545

(v) COMPUTER READABLE FORM:
        (A) MEDIUM TYPE: 3.5 inch diskette
        (B) COMPUTER: IBM compatible
        (C) OPERATING SYSTEM: Windows 3.1
        (D) SOFTWARE: Microsoft Word 5.1

(vi) CURRENT APPLICATION DATA:
        (A) APPLICATION NUMBER:
        (B) FILING DATE:
        (C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
        (A) APPLICATION NUMBER: PCT/US96/09513
        (B) FILING DATE: 06-JUN-96

(vii) PRIOR APPLICATION DATA:
        (A) APPLICATION NUMBER: PCT/US95/12791
        (B) FILING DATE: 12-OCT-95

(viii) ATTORNEY/AGENT INFORMATION:
        (A) NAME: Stephen C. Macevicz
        (B) REGISTRATION NUMBER: 30,285
        (C) REFERENCE/DOCKET NUMBER: 813wo

(ix) TELECOMMUNICATION INFORMATION:
        (A) TELEPHONE: (510) 670-9365
        (B) TELEFAX: (510) 670-9302

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 11 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

CTAGTCGACC A   11

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 11 nucleotides
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

NRRGATCYNN N   11

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 38 nucleotides
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA   38

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 20 nucleotides
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: double
      (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG   20

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 20 nucleotides
      (B) TYPE: nucleic acid

        (C) STRANDEDNESS: double
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GGGGCCCAGT CAGCGTCGAT   20

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 20 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: single
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATCGACGCTG ACTGGGCCCC   16

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
        (A) LENGTH: 62 nucleotides
        (B) TYPE: nucleic acid
        (C) STRANDEDNESS: double
        (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGG ATCCTCTTCC   50
TCTTCCTCTT CC   62

I claim:

1.   A method of determining the toxicity of a compound, the method comprising
the steps of:

5   administering the compound to a test organism;

           extracting a population of mRNA molecules from each of one or more tissues
of the test organism;

           forming a separate population of cDNA molecules from each population of
mRNA molecules from the one or more tissues such that each cDNA molecule of a

10   separate population has an oligonucleotide tag attached, the oligonucleotide tags
being selected from the same minimally cross-hybridizing set;

           separately sampling each population of cDNA molecules such that
substantially all different cDNA molecules within a separate population have different
oligonucleotide tags attached;

15   sorting the cDNA molecules of each separate population by specifically
hybridizing the oligonucleotide tags with their respective complements, the respective
complements being attached as uniform populations of substantially identical
complements in spatially discrete regions on one or more solid phase supports;

           determining the nucleotide sequence of a portion of each of the sorted cDNA

20   molecules of each separate population to form a frequency distribution of expressed
genes for each of the one or more tissues; and

           correlating the frequency distribution of expressed genes in each of the one or
more tissues with the toxicity of the compound.

25   2.   The method of claim 1 wherein said oligonucleotide tag and said complement
of said oligonucleotide tag are single stranded.

3.   The method of claim 2 wherein said oligonucleotide tag consists of a plurality
of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in

30   length and each subunit being selected from the same minimally cross-hybridizing set.

4.   The method of claim 3 wherein said one or more solid phase supports are
microparticles and wherein said step of sorting said cDNA molecules onto the
microparticles produces a subpopulation of loaded microparticles and a subpopulation

35   of unloaded microparticles.

5.   The method of claim 4 further including a step of separating said loaded
microparticles from said unloaded microparticles.

6.   The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

5

7.   The method of claim 6 wherein said number of loaded microparticles is at least 100,000.

8.   The method of claim 7 wherein said number of loaded microparticles is at
10   least 500,000.

9.   The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule
15   present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

10.   The method of claim 4 wherein said test organism is a mammalian tissue
- culture.
20

11.   The method of claim 10 wherein said mammalian tissue culture comprises hepatocytes.

12.   The method of claim 4 wherein said test organism is an animal selected from
25   the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

13.   The method of claim 12 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large
30   intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

14.   A method of identifying genes which are differentially expressed in a selected
35   tissue of a test animal after treatment with a compound, the method comprising the steps of:
        administering the compound to a test animal;

extracting a population of mRNA molecules from the selected tissue of the
test animal;

forming a population of cDNA molecules from the population of mRNA
molecules such that each cDNA molecule has an oligonucleotide tag attached, the
5   oligonucleotide tags being selected from the same minimally cross-hybridizing set;

sampling the population of cDNA molecules such that substantially all
different cDNA molecules have different oligonucleotide tags attached;

sorting the cDNA molecules by specifically hybridizing the oligonucleotide
tags with their respective complements, the respective complements being attached as
10   uniform populations of substantially identical complements in spatially discrete
regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA
molecules to form a frequency distribution of expressed genes; and

identifying genes expressed in response to administering the compound by
15   comparing the frequencing distribution of expressed genes of the selected tissue of the
test animal with a frequency distribution of expressed genes of the selected tissue of a
control animal.

15.   The method of claim 14 wherein said oligonucleotide tag and said
20   complement of said oligonucleotide tag are single stranded.

16.   The method of claim 15 wherein said oligonucleotide tag consists of a
plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9
nucleotides in length and each subunit being selected from the same minimally cross-
25   hybridizing set.

17.   The method of claim 16 wherein said one or more solid phase supports are
microparticles and wherein said step of sorting said cDNA molecules onto the
microparticles produces a subpopulation of loaded microparticles and a subpopulation
30   of unloaded microparticles.

18.   The method of claim 17 further including a step of separating said loaded
microparticles from said unloaded microparticles.

35   19.   The method of claim 18 further including a step of repeating said steps of
sampling, sorting, and separating until a number of said loaded microparticles is
accumulated is at least 10,000.

20.   The method of claim 19 wherein said number of loaded microparticles is at least 100,000.

21.   The method of claim 20 wherein said number of loaded microparticles is at least 500,000.

22.   The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

23.   The method of claim 17 wherein said test animal is selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

24.   The method of claim 23 wherein said selected tissue is selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

25.   A use of the technique of massively parallel signature sequencing to determine the toxicity of a compound in a test organism, the use comprising the steps of:
        administering the compound to a test organism;
        extracting a population of mRNA molecules from each of one or more tissues of the test organism and forming a population of cDNA molecules for each of the one or more tissues;
        determining the nucleotide sequence of a portion of each of the cDNA molecules of each separate population using massively parallel signature sequencing to form a frequency distribution of expressed genes for each of the one or more tissues; and
        correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

26.   The use of claim 25 wherein said test organism is a mammalian tissue culture.

27.   The use of claim 26 wherein said mammalian tissue culture comprises hepatocytes.

28.   The use of claim 25 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

5   29.   · The use of claim 28 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

10   30.   A use of the technique of massively parallel signature sequencing to identify genes which are differentially expressed in a test organism after treatment with a compound and which are correlated with toxicity of the compound, the use comprising the steps of:
            administering the compound to the test organism;
15   extracting a population of mRNA molecules from a selected tissue of the test organism and forming a population of cDNA molecules;
            determining the nucleotide sequence of a portion of each of the cDNA molecules using massively parallel signature sequencing to form a frequency
    - distribution of expressed genes;
20   identifying genes expressed in response to administering the compound by comparing the frequencing distribution of expressed genes of the selected tissue of the test organism with a frequency distribution of expressed genes of the selected tissue of a control organism; and
            determining whether the genes expressed in response to administering the
25   compound are correlated with toxicity of the compound in the test organism.

1/2

Fig. 1

Fig. 2

INTERNATIONAL SEARCH REPORT

A.	CLASSIFICATION OF SUBJECT MATTER

IPC(6)   : C12Q 1/68; C07H 21/04
US CL   : 435/6; 536/24.3
According to International Patent Classification (IPC) or to both national classification and IPC

B.	FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

  U.S. :   435/6; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

  APS, MEDLINE, BIOSIS, CAPLUS, SCISEARCH
  search terms: Martin, David W., toxic?, differential?, express?, cDNA, mRNA, RNA, gene#, hybrid?,

C.	DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHETVERIN et al. Oligonucleotide arrays: New concepts and possibilities. Bio/Technology. 12 November 1994, Vol. 12, pages 1093-1099, especially pages 1095-1096.	1-30
A	BRENNER et al. Encoded combinatorial chemistry. Proceedings of the National Academy of Sciences USA. June 1992, Vol. 89, pages 5381-5383.	1-30
A	MATSUBARA et al. cDNA analyses in the human genome project. Gene. 15 December 1993, Vol. 135, No. 1-2, pages 265-274.	1-30

☒   Further documents are listed in the continuation of Box C.   ☐   See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)		
		"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 JANUARY 1997	**19 FEB 1997**

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	SCOTT D. PRIEBE
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, page 4, lines 1-4, page 5, lines 31-37, page 17, lines 15-27, page 18, lines 30-35, page 20, line 23 to page 21, line 4.	1-30

FOCUS - 17 of 19 DOCUMENTS

August 11, 1997, Monday

SECTION: Financial News

DISTRIBUTION: TO BUSINESS AND MEDICAL EDITORS

LENGTH: 478 words

HEADLINE: Eli Lilly & Co. and **Acacia Biosciences** Enter Into Research Collaboration;
First Corporate Agreement for Acacia's Genome Reporter Matrix(TM)

DATELINE: RICHMOND, Calif., Aug. 11

BODY:

Acacia Biosciences and Eli Lilly and Company (Lilly) announced today the signing of a joint research collaboration to utilize Acacia's Genome Reporter Matrix(TM) (GRM) to aid in the selection and optimization of lead compounds. Under the collaboration, Acacia will provide chemical and biological profiles on a class of Lilly's compounds for an undisclosed fee.

Acacia's GRM is an assay-based computer modeling system that uses yeast as a miniature ecosystem. The GRM can profile the extent, nature and quantity of any changes in gene expression. Because of the similarities between the yeast and human genome, the system serves as an excellent surrogate for the human body, mimicking the effects induced by a biologically active molecule.

"Using yeast as a model organism for lead optimization makes a lot of sense given the high degree of homology with human metabolic pathways," said William Current of Lilly Research Laboratories. "Acacia's innovative GRM has the potential to provide enormous insight into the therapeutic impact of our compounds and make the drug discovery process more rational. It should substantially accelerate the development process."

"This first agreement with a major pharmaceutical company is an important milestone in the development of Acacia," said Bruce Cohen, President and CEO of Acacia. "The deal is in line with our strategy of establishing alliances that will allow our collaborators to use genomic profiles to identify and optimize compounds within their existing portfolios. In the long run, this technology can be used to characterize large scale combinatorial libraries, predict side effects prior to clinical trials and resurrect drugs that have failed during clinical trials."

The GRM incorporates two critical elements: chemical response profiles and genetic response profiles. The chemical response profiles measure the change in gene expression caused by potential therapeutics and then rank genes with altered expressions by degree of response. The genetic response profiles measure changes in gene expression caused by mutations in the genes encoding potential targets of pharmaceuticals; these genetic response profiles represent gold standards in drug discovery by defining the response profile expected for drugs with perfect selectivity and specificity. By comparing the two profiles, one can analyze a potential drug candidate's ability to mimic the action of a 'perfect' drug.

Acacia Biosciences is a functional genomics company developing proprietary technologies to enhance the speed and efficacy of drug discovery and development. Acacia's Genome Reporter Matrix capitalizes on the latest advances in genomics and combinatorial chemistry to generate comprehensive profiles of drug candidates' in vivo activity.
SOURCE Acacia Biosciences
CONTACT: Bruce Cohen, President and CEO of Acacia Biosciences, 510-669-2330 ext. 103 or Media: Linda Seaton of Feinstein

LOAD-DATE: August 12, 1997

GENETIC ENGINEERING NEWS

GEN

BIOTECHNOLOGY • BIOPROCESS • BIORESEARCH • TECHNOLOGY TRANSFER

09/25/97

Pharmagene Raises More Capital for Research on Human Tissues

By Sophia Fox

Pharmagene, the Royston, U.K.-based biopharmaceutical company specialising in the use of human biomaterials for drug discovery research, has raised a further £5 million from a group of investors led by 3i and Abacus Nominees. The funding will enable the company to expand both its human biomaterials collection and its capabilities across a range of proprietary platform technologies.

Gordon Baxter, Ph.D., Pharmagene's cofounder and chief operating officer, claimed, "by the end of this year Pharmagene will have access to the largest collection of human RNAs and proteins anywhere in the world, and a range of innovative, yet robust technologies

SEE PHARMAGENE, P. 9

Perkin-Elmer Acquires PerSeptive to Expand Its Capabilities in Gene-Based Drug Discovery

By John Sterling

Perkin-Elmer's (PE; Norwalk, CT) decision last month to acquire PerSeptive Biosystems (Framingham, MA) via a $360 million stock swap was designed to strengthen PE in terms of broad capabilities in gene-based drug discovery. The company's main goal is to develop new products to improve the integration of genetic and protein research.

"This merger will enhance our position as an effective provider of innovative, integrated platforms enabling our customers to be more efficient and cost-effective in bringing new pharmaceuticals to market," says Tony L. White, PE's chairman, president and CEO. "The combination of our two companies should bolster our presence in the life sciences, [and it is our] belief that we must take bold action now to lead the emerging era of molecular medicine with leading positions in both genetic and protein analysis."

A driving force behind the merger is the vast amount of genet-

ic information about human disease that is being accumulated by researchers and biotech companies working in the area of genomics. It is becoming increasingly obvious that these data need to be complemented with technologies for studying proteins and protein networks—a field known as proteomics (*see* GEN, *September 1, 1997, p.1*).

PE officials, who claim that MALDI-TOF (Matrix Assisted

SEE ACQUISITION, P. 10
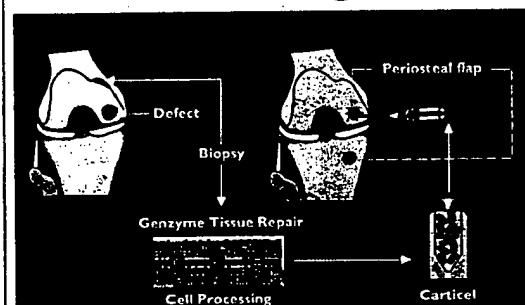
Perkin-Elmer acquired PerSeptive Biosystems for $360 million to obtain new technologies in mass spectrometry, bioseparations and purification for product development projects, spanning the range from genomics to proteomics.

FDA OKs Genzyme's Carticel Product for Damage to Knees

Carticel, which was approved for the repair of clinically significant, symptomatic cartilaginous defects of the femoral condyle (medial, lateral or trochlear) caused by acute or repetitive trauma, employs a proprietary process to grow autologous cartilage cells for implantation.

By Naomi Pfeiffer

The FDA has approved a knee-cartilage replacement product made by Genzyme Tissue Repair (Cambridge, MA), a tracking-stock division of Genzyme Corp., for people with trauma-damaged knees.

Carticel™ (autologous cultured chondrocytes) is the first product to be licensed under the FDA's pro-

SEE GENZYME, P. 6

Strategies for Target Validation Streamline Evaluation of Leads

By Vicki Glaser

Acacia Biosciences (Richmond, CA) last month announced its first agreement with a major pharmaceutical company, signing a deal with Eli Lilly (Indianapolis, IN) to use Acacia's Genome Reporter Matrix (GRM) to select and optimize some of Lilly's lead compounds. Acacia's yeast-based system for profiling drug activity is useful for evaluating the therapeutic potential of lead compounds, and it also has a role in the identification and validation of new drug targets.

"We're using the ecosystem of a cell to allow us to deduce the mechanism of action and target for any chemical," explains Bruce Cohen, president and CEO. "We screen for every target in a cell simultaneously...using transcription as a readout

for how a cell is adapting to any perturbation," he says.

The GRM technology consists of two main databases: one is the genetic response profile, showing the effects of mutations in each individual yeast gene and compensatory gene regulatory mechanisms; the other is the chemical response profile, which documents changes in gene expression in response to chemical compounds. Computational analysis and pattern matching between the genetic and chemical profiles yields information on the specificity, potency and side-effects risk of a drug lead.

Targeting Targets

No longer is mapping and sequencing a gene—or the human genome—an end unto itself, but

SEE TARGET, P. 15

Target

merely a means to an end. The critical next step is to validate the gene and its protein product as a potential drug target. The Human Genome Project continues to produce a treasure chest of expressed sequence tags (ESTs) and a tantalizing array of complete gene sequences.

Companies are applying a variety of functional genomic strategies to link genes to specific diseases and to multigenic phenotypes. Yet the ultimate challenge for pharmaceutical companies is to sift through all the sequence and differential gene expression data to identify the best targets for drug discovery.

Spinning off technology developed at the University of North Carolina (Chapel Hill), Cytogen Corp. (Princeton, NJ) formed its wholly owned subsidiary AxCell Biosciences earlier this year. The young company is building a protein interaction database, cataloging all the interactions the modular domains of proteins can engage in with a range of ligands, in order to gain insight into protein function and to select the most critical interaction to target for drug development.

AxCell's cloning-of-ligand-targets (COLT) technology employs "recognition units" from the company's genetic diversity library (GDL) to map functional protein interactions and quantitate their affinity. The company's inter-functional protcom-ic database (IFP-dbase) elucidates protein interaction networks and structure-activity relationships based on ligand affinity with protein modular domains.

Defining Disease Pathways

Signal Pharmaceuticals, Inc.'s (San Diego, CA) integrated drug target and discovery effort is based on mapping gene-regulating pathways in cells and identifying small molecules that regulate the activation of those genes. In collaboration with academic researchers, the company has identified a large number of regulatory proteins in several mitogen-activated protein (MAP) kinase pathways (including the JNK, FRK and p38

signaling pathways), which Signal is evaluating for the treatment of autoimmune, inflammatory, cardiovascular and neurologic diseases, and cancer. Other target identification

programs focus on the NF-kB pathway, estrogen-related genes and central/peripheral nervous system genes.

Regulating cytokine production in immune and inflammatory disorders,

The Genome Reporter Matrix depicts a subset of a yeast array. Each colony harbors a GFP-reporter construct for a single gene. Collectively, the array reports the expression of all yeast genes. A: Array in visible light. B: Image of fluorescent emission from the array.

Acacia
Biosciences

and modifying bone metabolism to treat osteoporosis are the focus of Signal's collaboration with Tanabe Seiyaku (Osaka, Japan). Signal has partnered with Organon/Akzo Nobel (Netherlands) to identify estrogen-responsive genes as targets for treating neurodegenerative and psychiatric diseases, atherosclerosis and ischemia, and with Roche Bioscience (Palo Alto, CA) to develop human peripheral nerve cell lines for the discovery of treatments for pain and incontinence.

Exelixis' (S. San Francisco, CA) strategy for target selection is to define disease pathways and identify regulatory molecules that activate or inhibit those biochemical/genetic pathways. Based on the finding that these pathways are conserved across species, the company is studying the model genetic systems of Drosophila and *Caenorhabditis elegans*. Using its PathFinder technology, Exelixis systematically introduces mutations into the genomes of these model organisms, looking for mutations that enhance or suppress the target disease-related gene. These novel genes then become the basis of drug screening assays.

Cadus Pharmaceutical Corp. (Tarrytown, NY) is identifying surrogate ligands to newly discovered orphan G-protein coupled transmembrane receptors of unknown function to determine the suitability of the receptors as drug targets. Inserting the novel receptor in a yeast system yields a ligand that activates the receptor. Access to a surrogate ligand allows the company to screen for receptor antagonists in the yeast system.

"The antagonist plus the surrogate ligand gives you two probes—an on probe and an off probe—which allows you to look at function," explains David Webb, Ph.D., vp of research and chief scientific officer. A surrogate ligand also provides information on which G-protein interacts with the orphan receptor and its associated signaling pathways, further clarifying the role of the receptor as a potential drug target. Cadus' collaboration with SmithKline (Philadelphia) capitalizes on Cadus' ability to determine orphan receptor function, applying the technology to SmithKline's proprietary, newly discovered G-protein receptors.

Cadus' recombinant yeast system can also be used to screen cell and tissue extracts for natural ligands, and the company is accelerating its internal drug-discovery efforts in the areas of cancer, inflammation and allergy. A recent equity investment in **Axiom Biotechnologies** (San Diego, CA) gave Cadus a license to Axiom's high-throughput pharmacologic screening system for lead optimization and discovery.

As its name implies, **gene/Networks** (Alameda, CA) focuses on identifying gene networks that contribute to multigenic phenotypes and complex disease processes. The integration of mouse and human genetic studies forms the basis of the technology. The Genome Tagged Mice database in development will serve as a library of natural mouse genetic and phenotypic variation. Disease-related genes identified in mice are then evaluated in human family- and population-based studies to confirm their clinical relevance and linkages to pathophysiologic traits.

Blocking Gene Expression

Inactivating a gene known to be expressed in association with a particular disease is one approach to identifying appropriate therapeutic targets. The target validation and discovery program at **Ribozyme Pharmaceuticals, Inc.** (Boulder, CO) applies the company's ribozyme technology to achieve selective inhibition of gene expression in cell culture and in animals.

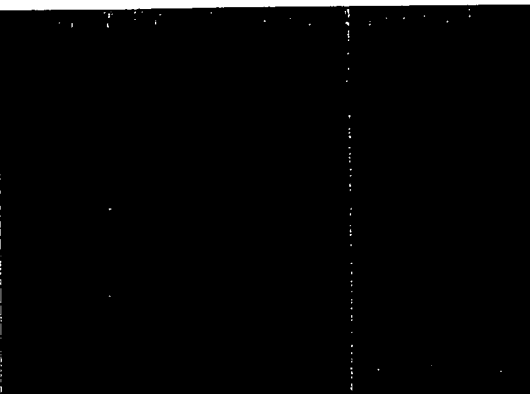Correlation of the gene expression inhibition with phenotype can

---

Target

suggest the relative importance of the gene in disease pathology. The company's nuclease-resistant ribozymes form the basis of a collaboration with Schering AG (Germany) for drug target validation and the development of ribozyme-based therapeutic agents, and with Chiron Corp. (Emeryville, CA) for target validation.

With several antisense compounds now progressing through clinical trials, the concept of using oligonucleotides to inhibit gene activity is not new. But rather than focusing on therapeutics development, Sequitur, Inc. (Natick, MA) is creating antisense compounds for the purpose of determining gene function and validating drug targets. Clients typically provide the one-year-old company with the sequence (or EST) of a potential gene target and, in return, Sequitur custom designs a series of three to six antisense compounds that yield a three-to-ten-fold inhibition of the target gene in cell culture. The company also provides oligofectins, a series of cationic lipids, to deliver the oligonucleotides to a variety of cultured cells.

"Differential expression information is just for correlation, it doesn't tell function or confirm what would be a good target," says Tod Woolf, Ph.D., director of technology development at Sequitur. Whereas, antisense compounds will inhibit a target. Sequitur offers both phosphorothioate DNA antisense compounds, and its proprietary Next Generation chimeric oligonucleotides, which have a higher hybridization affinity, greater specificity and reduced toxicity, according to the company.

Mining Pathogen Genomes

Companies such as Human Genome Sciences (HGS; Rockville, MD), Incyte (Palo Alto, CA),

Millennium Pharmaceuticals Inc. (Cambridge, MA) and Genome Therapeutics (Waltham, MA) are relying on high-speed DNA sequencing, positional cloning and other strategies to identify specific microbial genomic sites that would be good targets for infectious disease therapeutics.

HGS recently completed sequencing of the bacterial pathogen *Streptococcus pneumoniae*, which is the focus of an agreement with Hoffmann-La Roche (Basel, Switzerland). Roche will use the sequence data to develop new anti-infectives against *S. pneumoniae*. HGS and Roche have expanded their collaboration to include a nonexclusive license to access sequence information for the intestinal bacterium *Enterococcus faecalis*.

Incyte Pharmaceuticals has completed one-fold coverage of the *Candida albicans* genome, identify-

ing 60% of the genes of this fungal pathogen. This genome will become part of the company's PathoSeq microbial database. Incyte recently introduced the ZooSeq animal gene sequence and expression database. The database will provide genomic information across various species commonly used in preclinical drug testing, which may help to better define potential drug targets.

Millennium Pharmaceuticals continues to report success in identifying novel drug targets, having recently discovered a novel chemokine called neurotactin and a new class of MAD-related proteins that inhibit transforming growth factor beta (TGF-ß) signaling. The company also received U.S. patent coverage for the *tub* genes, believed to play a role in obesity, and for the gene that encodes the protein melastatin, which appears to suppress metastasis in malignant melanoma.　■

AxCell Biosciences scientists say their technology enables the rapid and simple functional identification of the two essential molecular components of protein interaction networks: specific recognition units that bind distinct modular protein domains are identified and isolated using a combination structural/functional approach that uses both peptide phase display Genetic Diversity Libraries (GDL) and bioinformatics, and cloning of Ligand Targets (COLT) technology utilizes recognition units as functional probes to isolate families of interactor proteins.

Pangea

Smith, now a computer programmer, is an expert in systems integration, Internet technologies and the application of industrial engineering principles to the drug discovery process. Before co-founding Pangea, he was the manager of software development at Attorney's Briefcase, a legal research software company.

By being "in the trenches" with customers and collaborators, Bellenson and Smith sensed the frustration of pharmaceutical researchers whose incompatible tools have impeded their progress. According to Bellenson, "Most of them are geared toward analyzing one molecule at a time. It's like emptying the ocean with an eye dropper—an incompatible eye dropper at that. A pharmaceutical company may have 30 different drug discovery teams with various approaches. The problem is to manage the process of experimenting with a lot of different approaches, to automate while maintaining flexibility."

GeneWorld 2.1 enables "integration of the entire target discovery and validation process," Bellenson says. The commercial software package coordinates the entire process of sequence-data analysis and can be integrated with other programs and databases, according to Smith, who adds that it handles thousands of sequence results, organizes and automates annotation and seamlessly interacts with growing genome databases. Simple forms and menus enable users to turn raw sequence data into crucial knowledge for drug discovery by applying algorithms to sequences, creating custom analysis strategies and producing useful reports, without the need for writing computer code. GeneWorld 2.1 runs on a variety of platforms and operating systems.

Pairing industrial relational database-management systems with a web-browser interface, Pangea's Operating System of Drug Discovery" is an open-computing framework that allows client/server and Java-enabled web-based technologies to collect, organize and analyze drug discovery information for pharmaceutical companies to simplify and accelerate drug discovery. The technology unites automated genomics database analysis for drug target site selection, chemical information database analysis and large-scale combinatorial chemistry project management and high-throughput screening project management for drug lead efficacy analysis. Pangea officials maintain that these integrated elements provide a unified environment for chemists, biologists and others involved in the drug discovery process to work together with

commercial and public domain software.

Pangea's Operating System of Drug Discovery can accommodate Sybase, Oracle or Informix relational database-management systems and any version of UNIX. It absorbs new data formats, databases, algorithms and analysis paradigms into the automated workflow without software modifications. Netscape Navigator" provides a friendly user interface from PC, Macintosh, and UNIX workstations.

In the near term, Pangea plans to complete its bioinformatics core with two more programs. Gene Foundry, a sample tracking and workflow sequence package for DNA sequence and fragment information, will also offer interaction with robots, reagent tracking and troubleshooting. Gene Thesaurus, the other package is a "warehouse of bioinformatics data," says Bellenson.　■

HIGH SPECIFIC ACTIVITY MICROBIAL ALKALINE PHOSPHATASE from Biocatalysts

---

Europe

GTAC Chairman, Professor Norman C. Nevin, said 1996 saw "four important developments": an increase in enquiries and submissions made to GTAC; an increase in the complexity of submitted protocols; a continuing shift from gene therapy for single-gene disorders toward strategies aimed at tumour destruction in cancer; and a growth in international sponsorship of U.K. gene therapy trials.

Since 1993, GTAC and its predecessor, the Clothier Committee, have approved 18 U.K. gene therapy clinical trials (13 of which have been carried out), which are listed in the report. The disease areas targeted by these trials include severe combined immunodeficiency (1 trial), cystic fibrosis (6), metastatic melanoma (2), lymphoma (2), neuroblastoma (1), breast cancer (1), Hurler's syndrome (1), cervical cancer (1), glioblastoma

breast cancer, breast cancer with liver metastases, glioblastoma, malignant ascites due to gastrointestinal cancer and ovarian cancer.

Copies of the GTAC third annual report are available from the GTAC Secretariat, Wellington House, 133-155 Waterloo Road, London SE1 8UG, U.K.

Coated Lenses Prevent PCO

Scientists in the U.K. say it may be possible to prevent posterior capsule opacification (PCO), a common complication following cataract surgery, by using the implanted poly-methylmethacrylate (PMMA) intraocular lens as a drug delivery system. PCO occurs in 30-50% of cataract surgery patients as a result of stimulated cell growth within the remaining capsular bag. The condition causes a decline in visual acuity and requires expensive laser treatment, thus negating the routine use of cataract surgery in underdeveloped countries, explains G. Duncan, at the

Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* 13, 3822 (1994); M. T. Madireddi *et al.*, *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E.

Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvarna, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho *et al.*, *Cell* 77, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1*, *2*), provide a practical and economical tool for studying gene expression on a very large scale (*3–6*).

Saccharomyces cerevisiae is an especially

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, *cis* regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

to any gene whose function is known (*15*). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl–coenzyme A(CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes PCK1, encoding phosphoenolpyruvate carboxykinase, and FBP1, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (*13*). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitochondrial biogenesis (*13*). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (*16–20*). A search in the promoter regions of the remaining two genes, ACR1 and IDP2, revealed that ACR1, a gene essential for ACS1 activity, also possessed a consensus CSRE motif, but interestingly, IDP2 did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception
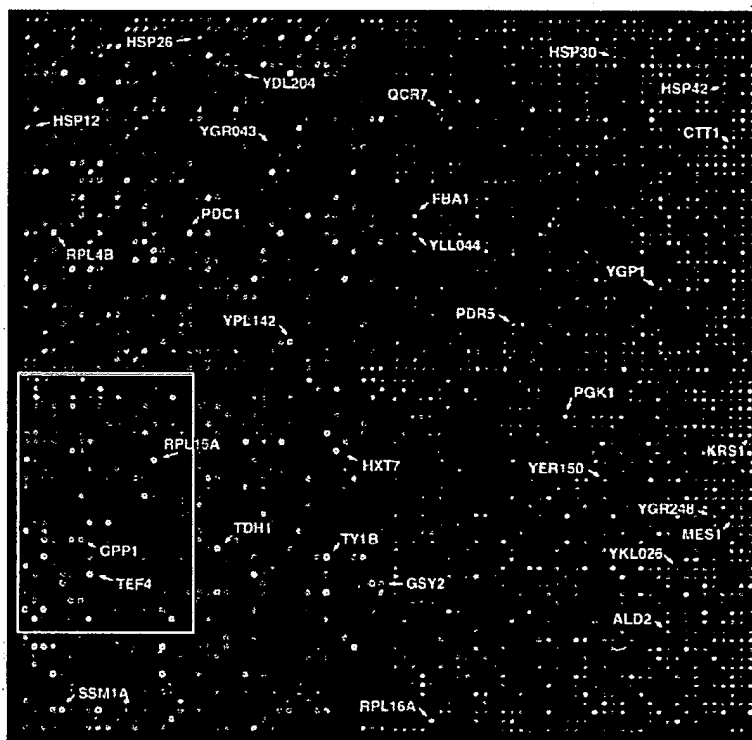

Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (*9*). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (*49*). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of <5 × 10⁶ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of ~2 × 10⁸ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of *HSP42*, have previously been shown to be controlled at least in part by these elements (*21–24*). Inspection of the sequences upstream of *HSP42* and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including *HSP30, ALD2, OM45,* and 10 uncharacterized ORFs (*25*)], nine contained one or more recognizable STRE sites in their upstream regions.

·The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (*26–28*). This complex binds a degenerate consensus sequence known as the CCAAT box (*26*). Computer analysis, using the consensus sequence TNRYTGGB (*29*), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (*30*). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of *HAP4* itself was induced nearly ninefold concurrent with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS$_{rpg}$) that is recognized by the Rap1 DNA-binding protein (*31, 32*). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (*33*). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (*34*). Indeed, we observed that the abundance of *RAP1* mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, *HAP4* and *SIP4*, were induced by a factor of more than threefold at the diauxic shift. *SIP4* encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (*35*). The eightfold induction of *SIP4* upon depletion of glucose strongly suggests a role in the induction of

downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (*36*). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (*37, 38*). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the *tup1Δ* mutation and *YAP1* overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (*13*). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (*39*). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (*40*).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*

Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a *MAT* α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1*Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a *MATA* strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, o-phenanthroline, heavy metals, and osmotic stress (*45*). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (*46, 47*). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (*48*). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.
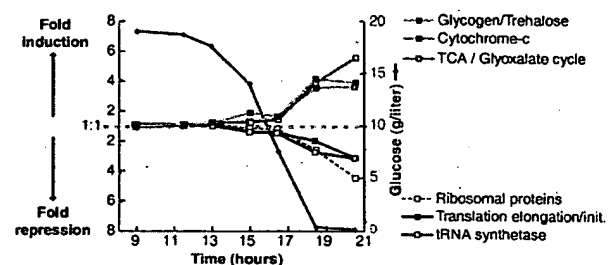

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (*50*). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Science 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, Genome Res. 6, 639 (1996).
3. D. Lashkari, Proc. Natl. Acad. Sci. U.S.A., in press.
4. J. DeRisi et al., Nature Genet. 14, 457 (1996).
5. D. J. Lockhart et al., Nature Biotechnol. 14, 1675 (1996).
6. M. Chee et al., Science 274, 610 (1996).
7. M. Johnston and M. Carlson, in The Molecular Biology of the Yeast Saccharomyces: Gene Expression, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-µl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-
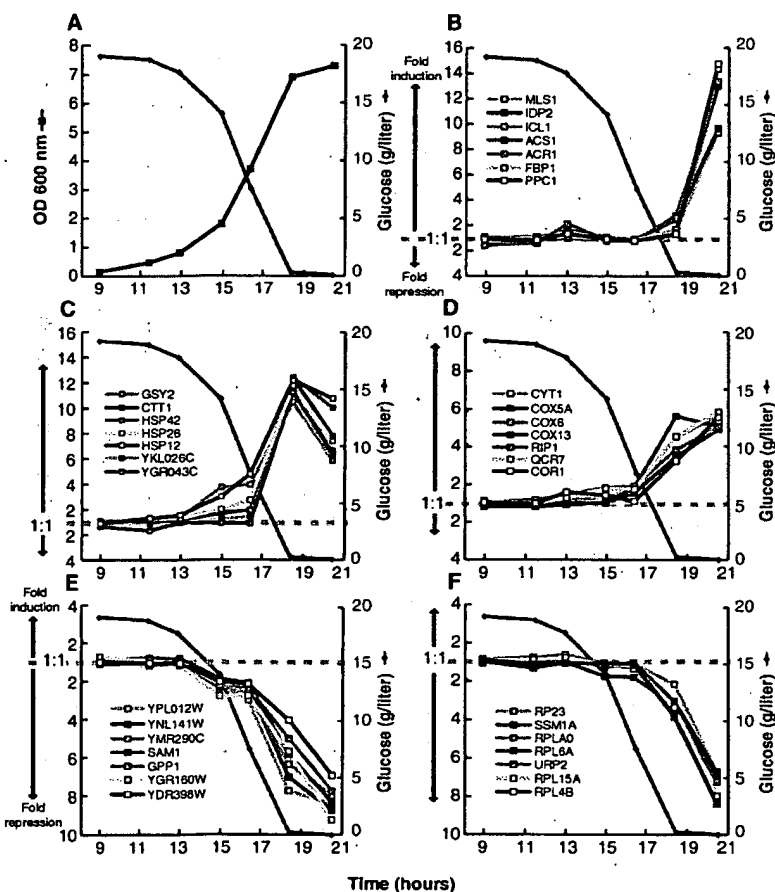
Time (hours)

Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of IDP2, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) SAM1, GPP1, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251) Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.

11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)+] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of .100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).

12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.

13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html

14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.

15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database (genome-www. stanford.edu), Yeast Protein Database (quest7. proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/ mips/yeast/index.htmlx).

16. A. Scholer and H. J. Schuller, Mol. Cell. Biol. 14, 3613 (1994).

17. S. Kratzer and H. J. Schuller, Gene 161, 75 (1995).

18. R. J. Haselbeck and H. L. McAlister, J. Biol. Chem. 268, 12116 (1993).

19. M. Fernandez, E. Fernandez, R. Rodicio, Mol. Gen. Genet. 242, 727 (1994).

20. A. Hartig et al., Nucleic Acids Res. 20, 5677 (1992).

21. P. M. Martinez et al., EMBO J. 15, 2227 (1996).

22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, Mol. Cell. Biol. 15, 6232 (1995).

23. H. Ruis and C. Schuller, Bioessays 17, 959 (1995).

24. J. L. Parrou, M. A. Teste, J. Francois, Microbiology 143, 1891 (1997).

25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.

26. S. L. Forsburg and L. Guarente, Genes Dev. 3, 1166 (1989).

27. J. T. Olesen and L. Guarente, ibid. 4, 1714 (1990).

28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, Mol. Microbiol. 13, 119 (1994).

29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B–C, G, or T; N–G, A, T, or C; R–A or G; and Y–C or T.

30. C. Fondrat and A. Kalogeropoulos, Comput. Appl. Biosci. 12, 363 (1996).

31. D. Shore, Trends Genet. 10, 408 (1994).

32. R. J. Planta and H. A. Raue, ibid. 4, 64 (1988).

33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.

34. S. F. Neuman, S. Bhattacharya, J. R. Broach, Mol. Cell. Biol. 15, 3187 (1995).

35. P. Lesage, X. Yang, M. Carlson, ibid. 16, 1921 (1996).

36. For example, we observed large inductions of the genes coding for PCK1, FBP1 [Z. Yin et al., Mol. Microbiol. 20, 751 (1996)], the central glyoxylate cycle gene ICL1 [A. Scholer and H. J. Schuller, Curr. Genet. 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, ACS1 [M. A. van den Berg et al., J. Biol. Chem. 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes PYK1 and PFK2 [P. A. Moore et al., Mol. Cell. Biol. 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T CTT1 [P. H. Bissinger et al., ibid. 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as HSP12, HSP26, and HSP42 [I. Farkas et al., J. Biol. Chem. 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, Mol. Gen. Genet. 223, 97 (1990); D. Wotton et al., J. Biol. Chem. 271, 2717 (1996)].

37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, FBP1 and PCK1) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).

38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, ADH1 and ADH2, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as HXK1/HXK2 (77% identical) [P. Herrero et al., Yeast 11, 137 (1995)], MLS1/ DAL7 (73% identical) (20), and PGM1/PGM2 (72% identical) [D. Oh, J. E. Hopper, Mol. Cell. Biol. 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.

39. F. E. Williams, U. Varanasi, R. J. Trumbly, Mol. Cell. Biol. 11, 3307 (1991).

40. D. Tzamarias and K. Struhl, Nature 369, 758 (1994).

41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concor-

dance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.

42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of TUP1 and an Eco R I site 124 base pairs before the stop codon of the TUP1 gene.

43. L. R. Kowalski, K. Kondo, M. Inouye, Mol. Microbiol. 15, 341 (1995).

44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, Gene 148, 149 (1994).

45. D. Hirata, K. Yano, T. Miyakawa, Mol. Gen. Genet. 242, 250 (1994).

46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, Appl. Environ. Microbiol. 60, 1783 (1994).

47. A. Muheim et al., Eur. J. Biochem. 195, 369 (1991).

48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, J. Biol. Chem. 269, 32592 (1994).

49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm. stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.

50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)

51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997

Response Under 37 C.F.R. 1.116 - Expedited Procedure
Examining Group 1646

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE
BEFORE THE BOARD OF PATENT APPEALS AND INTERFERENCES

In re Application of: Bandman et al.

Title: HUMAN GPCR PROTEINS

Serial No.:	09/895,686	Filing Date:	June 28, 2001
Examiner:	O'Hara, E	Group Art Unit:	1646

Mail Stop Appeal Brief-Patents
Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

BRIEF ON APPEAL

Sir:

Further to the Notice of Appeal filed November 5, 2003, and received by the USPTO on November 7, 2003, herewith are three copies of Appellants' Brief on Appeal. Authorized fees include the $ **330.00** fee for the filing of this Brief.

This is an appeal from the decision of the Examiner finally rejecting claims 1-6 of the above-identified application.

(1) REAL PARTY IN INTEREST

The above-identified application is assigned of record to Incyte Pharmaceuticals, Inc., (now Incyte Corporation, formerly known as Incyte Genomics, Inc.) (Reel 012272, Frame 0191) which is the real party in interest herein.

117940   09/895,686

(2) RELATED APPEALS AND INTERFERENCES

Appellants, their legal representative and the assignee are not aware of any related appeals or interferences which will directly affect or be directly affected by or have a bearing on the Board's decision in the instant appeal.

(3) STATUS OF THE CLAIMS

Claims rejected: Claims 1-6.

Claims allowed: (none).

Claims canceled: Claims 13-20.

Claims withdrawn: Claims 7-12.

Claims on Appeal: Claims 1-6 (A copy of the claims on appeal, as amended, can be found in the attached Appendix.)

(4) STATUS OF AMENDMENTS AFTER FINAL

There were no amendments submitted after Final Rejection.

(5) SUMMARY OF THE INVENTION

Appellants' invention is directed to polynucleotides encoding a human G-protein coupled receptor (GPCR), SEQ ID NO:1, in particular, a metabotropic glutamate GPCR, based on the conservation of various sequence motifs characteristic of this family of proteins, in particular, the seven hydrophobic transmembrane domains characteristic of GPCRs. See specification, at page 11 and Table 1 and Figure 1. The glutamate GPCRs are described in the specification and the art of record as important in neurotransmission and involved in neurological disorders such as epilepsy, stroke, and neurodegeneration. See specification, at page 2. Polynucleotides encoding SEQ ID NO:1 are also disclosed as differentially expressed in thyroid tumors, in particular, follicular carcinoma based on Northern analysis in thyroid tissues. See specification, at page 35. The claimed polynucleotides are asserted to be useful in the diagnosis, treatment, and evaluation of therapies for neurological and neoplastic disorders, in particular, follicular carcinoma.

(6) ISSUES

1.   Whether claims 1-6 directed to SEQ ID NO:1 encoding polynucleotides meet the utility requirement of 35 U.S.C. §101. In particular, whether the conservation of sequence motifs and domains between the protein coded for by the claimed polynucleotide and metabotropic GPCRs, known to have utility in neurotransmission and neurological disorders, demonstrates a "substantial likelihood" of utility under 35 U.S.C. § 101. Whether there is evidence that the differential expression of the polynucleotide encoding SEQ ID NO:1 in thyroid tumors provides a substantial likelihood of utility for the claimed polynucleotides in the detection and diagnosis of thyroid tumors.

2.   Whether one of ordinary skill in the art would know how to use the claimed polynucleotides, e.g., in toxicology testing, drug development, and the diagnosis of disease, so as to satisfy the enablement requirement of 35 U.S.C. §112, first paragraph.

3.   Whether fragments and variants of the polynucleotides encoding SEQ ID NO:1 are sufficiently described in the specification that the skilled artisan would recognize applicant's possession of them at the time the application was filed in accordance with 35 U.S.C. § 112, First Paragraph.

4.   Whether the claimed polynucleotides are sufficiently described in priority application Serial No. 09/516,513, filed September 17, 1998 to meet the requirements of 35 U.S.C. § 112, first paragraph and claim an effective priority date of September 17, 1998 with respect to the now claimed invention.

(7) GROUPING OF THE CLAIMS

As to Issue 1

All of the claims on appeal stand or fall together.

As to Issue 2

All of the claims on appeal stand or fall together

As to Issue 3

All of the claims on appeal stand or fall together

As to Issue 4

Claims 1 and 3-6 stand or fall together.

(8) APPELLANTS' ARGUMENTS

The rejection of claims 1-6 under 35 U.S.C. §§ 101 and 112, first paragraph is improper, as the inventions of those claims have a patentable utility as set forth in the instant specification, and/or a utility well known to one of ordinary skill in the art.

Claims 1-6 stand rejected under 35 U.S.C. §§ 101 and 112, first paragraph, based on the allegation that the claimed invention lacks patentable utility. The rejection alleges in particular that:

• the claimed invention is not supported by either a substantial and specific asserted utility or a well- established utility. None of the described uses are considered to be specific or substantial utilities for either the protein or encoding nucleic acid molecules. Methods such as identification of ligands, use to screen for homologous genes, use to identify chromosomes or chromosomal locations, use to recombinantly produce protein or to generate antibodies are considered general methods applicable to any protein and/or nucleic acid.

• Applicants assertion that the claimed polynucleotide can be used in cancer diagnosis, in particular follicular carcinoma of the thyroid, is unconvincing because the correlation between the expression of the polynucleotide and follicular carcinoma is based on one single library. The determination of a cancer marker must be based on studying results from considerable number of patients, and statistical analysis. See Guidelines for Marker Development by the National Cancer Institute (NCI).

The invention at issue is a polynucleotide corresponding to a gene that is expressed in humans. The novel polynucleotide codes for a polypeptide demonstrated in the patent specification to be a member of the class of glutmate GPCRs, whose biological functions include control of neuerotransmission. The claimed invention has numerous practical, beneficial uses in toxicology testing, drug development, and the diagnosis of disease, none of which requires

knowledge of how the polypeptide coded for by the polynucleotide actually functions. The claimed invention can be used, for example, as a marker for cancers of the thyroid, in particular, follicular carcinoma. See specification, at page 35. As a result of the benefits of these uses, the claimed invention already enjoys significant commercial success.

Applicants have previously submitted a Declaration by Dr. John C. Rockett showing the many reasons why the use of the claimed polynucleotides in gene expression profiling studies in toxicology testing would be readily apparent to the skilled artisan at the time the application was filed.

Applicants further submit two additional expert Declarations by Dr. Vishwanath R. Iyer and Dr. Tod Bedilion under 37 C.F.R. § 1.132, with respective attachments, and ten (10) scientific references filed before the September 17, 1998 priority date of the instant application. The Rockett Declaration, Iyer Declaration, Bedilion Declaration, and the ten (10) references fully establish that, prior to the September 17, 1998 filing date of the parent Bandman '513 application, it was well-established in the art that:

polynucleotides derived from nucleic acids expressed in one or more tissues and/or cell types can be used as hybridization probes -- that is, as tools -- to survey for and to measure the presence, the absence, and the amount of expression of their cognate gene;

with sufficient length, at sufficient hybridization stringency, and with sufficient wash stringency -- conditions that can be routinely established -- expressed polynucleotides, used as probes, generate a signal that is specific to the cognate gene, that is, produce a gene-specific expression signal;

expression analysis is useful, *inter alia*, in drug discovery and lead optimization efforts, in toxicology, particularly toxicology studies conducted early in drug development efforts, and in phenotypic characterization and categorization of cell types, including neoplastic cell types;

each additional gene-specific probe used as a tool in expression analysis provides an additional gene-specific signal that could not otherwise have been detected, giving a more comprehensive, robust, higher resolution, statistically more significant, and thus more useful expression pattern in such analyses than would otherwise have been possible;

biologists, such as toxicologists, recognize the increased utility of more comprehensive, robust, higher resolution, statistically more significant

results, and thus want each newly identified expressed gene to be included in such an analysis;

nucleic acid microarrays increase the parallelism of expression measurements, providing expression data analogous to that provided by older, lower throughput techniques, but at substantially increased throughput;

accordingly, when expression profiling is performed using microarrays, each additional gene-specific probe that is included as a signaling component on this analytical device increases the detection range, and thus versatility, of this research tool;

biologists, such as toxicologists, recognize the increased utility of such improved tools, and thus want a gene-specific probe to each newly identified expressed gene to be included in such an analytical device;

the industrial suppliers of microarrays recognize the increased utility of such improved tools to their customers, and thus strive to improve salability of their microarrays by adding each newly identified expressed gene to the microarrays they sell;

it is not necessary that the biological function of a gene be known for measurement of its expression to be useful in drug discovery and lead optimization analyses, toxicology, or molecular phenotyping experiments;

failure of a probe to detect changes in expression of its cognate gene does not diminish the usefulness of the probe as a research tool; and

failure of a probe completely to detect its cognate transcript in any single expression analysis experiment does not deprive the probe of usefulness to the community of users who would use it as a research tool.]

The Patent Examiner does not dispute that the claimed polynucleotide can be used as a probe in cDNA microarrays and used in gene expression monitoring applications. Instead, the Patent Examiner contends that the claimed polynucleotide cannot be useful without precise knowledge of its biological function, or the biological function of the polypeptide it encodes. But the law has never required knowledge of biological function to prove utility. It is the claimed invention's uses, not its functions, that are the subject of a proper analysis under the utility requirement.

In any event, as demonstrated by the Rockett Declaration, the Iyer Declaration, and the Bedilion Declaration, the person of ordinary skill in the art can achieve beneficial results from the claimed polynucleotide in the absence of any knowledge as to the precise function of the

protein encoded by it. The uses of the claimed polynucleotide in gene expression monitoring applications are in fact independent of its precise biological function.

I.   The applicable legal standard

To meet the utility requirement of sections 101 and 112 of the Patent Act, the patent applicant need only show that the claimed invention is "practically useful," *Anderson v. Natta*, 480 F.2d 1392, 1397, 178 USPQ 458 (CCPA 1973) and confers a "specific benefit" on the public. *Brenner v. Manson*, 383 U.S. 519, 534-35, 148 USPQ 689 (1966). As discussed in a recent Court of Appeals for the Federal Circuit case, this threshold is not high:

> An invention is "useful" under section 101 if it is capable of providing some identifiable benefit. See *Brenner v. Manson*, 383 U.S. 519, 534 [148 USPQ 689] (1966); *Brooktree Corp. v. Advanced Micro Devices, Inc.*, 977 F.2d 1555, 1571 [24 USPQ2d 1401] (Fed. Cir. 1992) ("to violate Section 101 the claimed device must be totally incapable of achieving a useful result"); *Fuller v. Berger*, 120 F. 274, 275 (7th Cir. 1903) (test for utility is whether invention "is incapable of serving any beneficial end").

Juicy Whip Inc. v. Orange Bang Inc., 51 USPQ2d 1700 (Fed. Cir. 1999).

While an asserted utility must be described with specificity, the patent applicant need not demonstrate utility to a certainty. In *Stiftung v. Renishaw PLC*, 945 F.2d 1173, 1180, 20 USPQ2d 1094 (Fed. Cir. 1991), the United States Court of Appeals for the Federal Circuit explained:

> An invention need not be the best or only way to accomplish a certain result, and it need only be useful to some extent and in certain applications: "[T]he fact that an invention has only limited utility and is only operable in certain applications is not grounds for finding lack of utility." *Envirotech Corp. v. Al George, Inc.*, 730 F.2d 753, 762, 221 USPQ 473, 480 (Fed. Cir. 1984).

The specificity requirement is not, therefore, an onerous one. If the asserted utility is described so that a person of ordinary skill in the art would understand how to use the claimed invention, it is sufficiently specific. *See Standard Oil Co. v. Montedison, S.p.a.*, 212 U.S.P.Q. 327, 343 (3d Cir. 1981). The specificity requirement is met unless the asserted utility amounts to a "nebulous expression" such as "biological activity" or "biological properties" that does not convey meaningful information about the utility of what is being claimed. *Cross v. Iizuka*, 753 F.2d 1040, 1048 (Fed. Cir. 1985).

In addition to conferring a specific benefit on the public, the benefit must also be "substantial." *Brenner*, 383 U.S. at 534. A "substantial" utility is a practical, "real-world" utility. *Nelson v. Bowler*, 626 F.2d 853, 856, 206 USPQ 881 (CCPA 1980).

If persons of ordinary skill in the art would understand that there is a "well-established" utility for the claimed invention, the threshold is met automatically and the applicant need not make any showing to demonstrate utility. Manual of Patent Examining Procedure at § 706.03(a). Only if there is no "well-established" utility for the claimed invention must the applicant demonstrate the practical benefits of the invention. *Id.*

Once the patent applicant identifies a specific utility, the claimed invention is presumed to possess it. *In re Cortright*, 165 F.3d 1353, 1357, 49 USPQ2d 1464 (Fed. Cir. 1999); *In re Brana*, 51 F.3d 1560, 1566; 34 USPQ2d 1436 (Fed. Cir. 1995). In that case, the Patent Office bears the burden of demonstrating that a person of ordinary skill in the art would reasonably doubt that the asserted utility could be achieved by the claimed invention. *Id.* To do so, the Patent Office must provide evidence or sound scientific reasoning. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). If and only if the Patent Office makes such a showing, the burden shifts to the applicant to provide rebuttal evidence that would convince the person of ordinary skill that there is sufficient proof of utility. *Brana*, 51 F.3d at 1566. The applicant need only prove a "substantial likelihood" of utility; certainty is not required. *Brenner*, 383 U.S. at 532.

II.   Toxicology Testing and disease diagnosis are sufficient utilities under 35 U.S.C. §§ 101 and 112, first paragraph

The claimed invention meets all of the necessary requirements for establishing a credible utility under the Patent Law: There are "well-established" uses for the claimed invention known to persons of ordinary skill in the art, and there are specific practical and beneficial uses for the invention disclosed in the patent application's specification. These uses are explained, in detail, in the Rockett Declaration, Iyer Declaration, and Second Bedilion Declaration accompanying this brief or previously submitted. Objective evidence, not considered by the Patent Office, further corroborates the credibility of the asserted utilities.

A.   The use of the claimed SEQ ID NO:1 encoding polynucleotides for toxicology
testing, drug discovery, and disease diagnosis are practical uses that confer
"specific benefits" to the public

The claimed invention has specific, substantial, real-world utility by virtue of its use in

toxicology testing, drug development and disease diagnosis through gene expression profiling.

These uses are explained in detail in the accompanying Rockett Declaration, Iyer Declaration,

and Bedilion Declaration, the substance of which is not rebutted by the Patent Examiner. There

is no dispute that the claimed invention is in fact a useful tool in cDNA microarrays used to

perform gene expression analysis. That is sufficient to establish utility for the claimed

polynucleotide.

In his Declaration, Dr. Rockett explains the many reasons why a person skilled in the art

in 1998 would have understood that any expressed polynucleotide is useful for a number of gene

expression monitoring applications, *e.g.*, in cDNA microarrays, in connection with the

development of drugs and the monitoring of the activity of such drugs. (Rockett Declaration at,

e.g., ¶¶ 10-18).

> It is my opinion, therefore, based on the state of the art in toxicology at least since
> the mid-1990s . . . that disclosure of the sequence of a new gene or protein, with or
> without knowledge of its biological function, would have been sufficient information
> for a toxicologist to use the gene and/or protein in expression profiling studies in
> toxicology.[1] [Rockett Declaration, ¶ 18.]

In his Declaration, Dr. Bedilion explains why a person of skill in the art in 1998 would

have understood that any expressed polynucleotide is useful for gene expression monitoring

applications using cDNA microarrays. (Bedilion Declaration, e.g., ¶¶ 4-7.) In his Declaration,

Dr. Iyer explains why a person of skill in the art in 1998 would have understood that any

expressed polynucleotide is useful for gene expression monitoring applications using cDNA

microarrays, stating that "[t]o provide maximum versatility as a research tool, the microarray

should include ☐ and as a biologist I would want my microarray to include ☐ each newly

identified gene as a probe." (Iyer Declaration, ¶ 9.)

"Use of the words 'it is my opinion' to preface what someone of ordinary skill in the art
would have known does not transform the factual statements contained in the declaration into
opinion testimony." *In re Alton*, 37 USPQ2d 1578, 1583 (Fed. Cir. 1996).

In addition, Dr. Rockett explains in his Declaration that "there are a number of other differential expression analysis technologies that precede the development of microarrays, some by decades, and that have been applied to drug metabolism and toxicology research, including: (1) differential screening; (2) subtractive hybridization, including variants such as chemical cross-linking subtraction, suppression-PCR subtractive hybridization and representational difference analysis; (3) differential display; (4) restriction endonuclease facilitated analyses, including serial analysis of gene expression (SAGE) and gene expression fingerprinting and (5) EST analysis." (Rockett Declaration, ¶ 7.)

Nowhere does the Patent Examiner address the fact that, as described on pages 31-32 of the Bandman '513 application, the claimed polynucleotides can be used as highly specific probes in, for example, cDNA microarrays - probes that without question can be used to measure both the existence and amount of complementary RNA sequences known to be the expression products of the claimed polynucleotides. The claimed invention is not, in that regard, some random sequence whose value as a probe is speculative or would require further research to determine.

Given the fact that the claimed polynucleotide is known to be expressed, its utility as a measuring and analyzing instrument for expression levels is as indisputable as a scale's utility for measuring weight. This use as a measuring tool, regardless of how the expression level data ultimately would be used by a person of ordinary skill in the art, by itself demonstrates that the claimed invention provides an identifiable, real-world benefit that meets the utility requirement. *Raytheon v. Roper*, 724 F.2d 951, (Fed. Cir. 1983) (claimed invention need only meet one of its stated objectives to be useful); *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999) (how the invention works is irrelevant to utility); MPEP § 2107 ("Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific, and unquestionable utility (e.g., they are useful in analyzing compounds)" (emphasis added)).

Literature reviews published shortly before the filing of the Bandman '513 application describing the state of the art further confirm the claimed invention's utility. Rockett et al. confirm, for example, that the claimed invention is useful for differential expression analysis regardless of how expression is regulated:

Despite the development of multiple technological advances which have recently brought the field of gene expression profiling to the forefront of molecular analysis, recognition of the importance of differential gene expression and characterization of differentially expressed genes has existed for many years.

* * *

Although differential expression technologies are applicable to a broad range of models, perhaps their most important advantage is that, in most cases, absolutely no prior knowledge of the specific genes which are up- or down-regulated is required.

* * *

Whereas it would be informative to know the identity and functionality of all genes up/down regulated by . . . toxicants, this would appear a longer term goal . . . . However, the current use of gene profiling yields a *pattern* of gene changes for a xenobiotic of unknown toxicity which may be matched to that of well characterized toxins, thus alerting the toxicologist to possible *in vivo* similarities between the unknown and the standard, thereby providing a platform for more extensive toxicological examination. (emphasis in original)

Rockett et al., <u>Differential gene expression in drug metabolism and toxicology: practicalities, problems and potential</u>, Xenobiotica 29:655-691 (July 1999) (Rockett Declaration, Exhibit C).

In another pre-September 1998 article, Lashkari et al. state explicitly that sequences that are merely "predicted" to be expressed (predicted Open Reading Frames, or ORFs)-- the claimed invention in fact is known to be expressed-- have numerous uses:

Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons-- they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. <u>The amplicons can also be used directly by, for example, arraying onto glass for expression analysis</u>, for DNA binding assays, or for any direct DNA assay. (emphasis added)

Lashkari et al., <u>Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR</u>, Proc. Nat. Acad. Sci. 94:8945-8947 (Aug. 1997) (Reference No. 1).

B. **The use of polynucleotides coding for polypeptides expressed by humans as tools for toxicology testing, drug discovery, and the diagnosis of disease is now "well-established"**

The technologies made possible by expression profiling and the DNA tools upon which they rely are now well-established. The technical literature recognizes not only the prevalence of these technologies, but also their unprecedented advantages in drug development, testing and safety assessment. These technologies include toxicology testing, e.g., as described by Bedilion, Rockett, and Iyer in their Declarations.

Toxicology testing is now standard practice in the pharmaceutical industry. See, *e.g.*, John C. Rockett et al., *supra*:

> Knowledge of toxin-dependent regulation in target tissues is not solely an academic pursuit as much interest has been generated in the pharmaceutical industry to harness this technology in the early identification of toxic drug candidates, thereby shortening the developmental process and contributing substantially to the safety assessment of new drugs. (Rockett Declaration, Exhibit C, page 656)

To the same effect are several other scientific publications, including Emile F. Nuwaysir et al., Microarrays and toxicology: The advent of toxicogenomics, Molecular Carcinogenesis 24:153-159 (1999) (Reference No. 2); Sandra Steiner and N. Leigh Anderson, Expression profiling in toxicology -- potentials and limitations, Toxicology Letters 112-13:467-471 (2000) (Reference No. 3).

Nucleic acids useful for measuring the expression of whole classes of genes are routinely incorporated for use in toxicology testing. Nuwaysir et al. describes, for example, a Human ToxChip comprising 2089 human clones, which were selected

> for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip.

See also Table 1 of Nuwaysir et al. (listing additional classes of genes deemed to be of special interest in making a human toxicology microarray).

The more genes that are available for use in toxicology testing, the more powerful the technique. "Arrays are at their most powerful when they contain the entire genome of the species they are being used to study." John C. Rockett and David J. Dix, Application of DNA arrays to toxicology, Environ. Health Perspec.107:681-685 (1999) (Reference No. 4). Control genes are carefully selected for their stability across a large set of array experiments in order to best study the effect of toxicological compounds. See attached email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding (Reference No. 5), indicating that even the expression of carefully selected control genes can be altered. Thus, there is no expressed gene which is irrelevant to screening for toxicological effects, and all expressed genes have a utility for toxicological screening.

Further evidence of the well-established utility of all expressed polypeptides and polynucleotides in toxicology testing is found in U.S. Pat. No. 5,569,588 (Reference No. 9e) and published PCT applications WO 95/21944 (Reference No. 9a), WO 95/20681 (Reference No. 9b), and WO 97/13877 (Reference No. 9g).

WO 95/21944 ("Differentially expressed genes in healthy and diseased subjects"), published August 17, 1995, describes the use of microarrays in expression profiling analyses, emphasizing that *patterns* of expression can be used to distinguish healthy tissues from diseased tissues and that *patterns* of expression can additionally be used in drug development and toxicology studies, without knowledge of the biological function of the encoded gene product. In particular, and with emphasis added:

> The present invention involves . . . methods for diagnosing diseases . . . characterized by the presence of [differentially expressed] . . . genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/ polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST. . . . Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of diseases based on differential expression of genes of unknown function . . . . [abstract]

> The method [of the present invention] involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA

polynucleotide sequences . . . and a defined set of oligonucleotide/polynucleotide[]
. . . immobilized on a support. Those defined [immobilized]
oligonucleotide/polynucleotide sequences are <u>representative of the total expressed
genetic component of the cells</u>, tissues, organs or organism as defined by the
collection of partial cDNA sequences (ESTs). [page 2]

The present invention meets the unfilled needs in the art by providing
methods for the . . . <u>use of gene fragments and genes, even those of unknown full
length sequence and unknown function, which are differentially expressed</u> in a
healthy animal and in an animal having a specific disease or infection by use of
ESTs derived from DNA libraries of healthy and/or diseased/infected animals.
[page 4]

Yet another aspect of the invention is that it provides . . . a means for . . .
monitoring the efficacy of disease treatment regimes <u>including . . . toxicological
effects thereof</u>." [page 4]

It has been appreciated that one or more differentially identified EST or
gene-specific oligonucleotide/polynucleotides <u>define a pattern</u> of differentially
expressed genes diagnostic of a predisease, disease or infective state. <u>A knowledge
of the specific biological function of the EST is not required</u> only that the EST[]
identifies a gene or genes whose altered expression is associated reproducibly with
the predisease, disease or infectious state. [page 4]

As used herein, the term 'disease' or 'disease state' refers to any condition
which deviates from a normal or standardized healthy state in an organism of the
same species in terms of differential expression of the organism's genes. . .
[whether] of genetic or environmental origin, for example, an inherited disorder
such as certain breast cancers. . . .[or] administration of a drug or exposure of the
animal to another agent, e.g., nutrition, which affects gene expression. [page 5]

As used herein, the term 'solid support' refers to any known substrate which
is useful for the immobilization of large numbers of oligonucleotide/polynucleotide
sequences by any available method . . . [and includes, inter alia,] nitrocellulose, . . .
glass, silica. . . . [page 6]

By 'EST' or 'Expressed Sequence Tag' is meant a partial DNA or cDNA
sequence of about 150 to 500, more preferably about 300, sequential nucleotides. . .
. [page 6]

One or more libraries made from a single tissue type typically provide at
least about 3000 different (i.e., unique) ESTs and <u>potentially the full complement of
all possible ESTs representing all cDNAs e.g., 50,000 to 100,000 in an animal such
as a human</u>. [page 7]

The lengths of the defined oligonucleotide/ polynucleotides may be readily increased or decreased as desired or needed. . . . The length is generally guided by the principle that it should be of sufficient length to insure that it is on[] average only represented once in the population to be examined. [page 7]

Comparing the . . . hybridization patterns permits detection of those defined oligonucleotide/ polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions [of the solid support]. [page 13]

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained[;] rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used [to] form diagnostic patterns or to identify which particular EST is detected. For example, all known ESTs from an organism are used to produce a 'master' solid support to which control sample and disease samples are alternately hybridized. [page 14]

Diagnosis is accomplished by comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This[,] like many of the foregoing embodiments[,] may use known or unknown ESTs derived from many libraries. [page 18]

Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal[s], especially humans. [page 18]

WO 95/20681 ("Comparative Gene Transcript Analysis"), filed in 1994 by Appellants' assignee and published August 3, 1995, has three issued U.S. counterparts: U.S. Pat. Nos. 5,840,484, issued November 24, 1998; 6,114,114, issued September 5, 2000; and 6,303,297, issued October 16, 2001.

The specification describes the use of transcript expression *patterns,* or "images", each comprising multiple pixels of gene-specific information, for diagnosis, for cellular phenotyping, and in toxicology and drug development efforts. The specification describes a plurality of methods for obtaining the requisite expression data -- one of which is microarray hybridization -- and equates the uses of the expression data from these disparate platforms. In particular, and with emphasis added:

The invention provides a "method and system for quantifying the relative abundance of gene transcripts in a biological specimen. . . . [G]ene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is <u>correlated</u> to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides <u>a method for comparing the gene transcript image analysis</u> from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens." [abstract]

"<u>[W]e see each individual gene product as a 'pixel' of information</u> which <u>relates to the expression of that, and only that, gene</u>. We teach herein [] methods whereby <u>the individual 'pixels' of gene expression information can be combined into a single gene transcript 'image,'</u>in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood." [page 2]

"The present invention avoids the drawbacks of the prior art by providing <u>a method to quantify the relative abundance of multiple gene transcripts in a given biological specimen</u>. . . . The method of the instant invention provides for detailed diagnostic <u>comparisons of cell profiles</u> revealing numerous changes in the expression of individual transcripts." [page 6]

"High resolution analysis of gene expression be <u>used directly as a diagnostic profile</u>. . . . " [page 7]

"The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed." [page 7]

"The invention . . . includes a method of comparing specimens containing gene transcripts." [page 7]

"The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens." [i.e., the results yield analogous data to microarrays] [page 8]

"Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made." [page 8]

"In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the differences and similarities." [page 9]

"In essence, the invention is a method and system for quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens. . . . " [page 9]

"[T]wo or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells." [pages 9 - 10]

"The present invention provides a method to compare the relative abundance of gene transcripts in different biological specimens. . . . This process is denoted herein as gene transcript imaging. The quantitative analysis of the relative abundance for a set of gene transcripts is denoted herein as 'gene transcript image analysis' or 'gene transcript frequency analysis'. The present invention allows one to <u>obtain a profile</u> for gene transcription <u>in any given population of cells or tissue from any type of organism</u>." [page 11]

"The invention has <u>significant advantages in the fields of diagnostics, toxicology and pharmacology,</u> to name a few." [page 12]

"[G]ene transcript sequence abundances are compared against reference database sequence abundances including normal data sets for diseased and healthy patients. The patient has the disease(s) with which the patient's <u>data set</u> most closely <u>correlates</u>." [page 12]

"For example, gene transcript frequency analysis can be used to differentiate normal cells or tissues from diseased cells or tissues. . . ." [page 12]

"<u>In toxicology</u>, . . . [g]ene transcript imaging provides highly detailed information on the cell and tissue environment, some of which would not be obvious in conventional, less detailed screening methods. <u>The gene transcript image is a more powerful method to predict drug toxicity and efficacy</u>. Similar benefits accrue in the use of this tool in pharmacology. . . . " [page 12]

"<u>In an alternative embodiment</u>, comparative gene transcript frequency analysis is <u>used to differentiate between cancer cells</u> which respond to anti-cancer agents and those which do not respond." [page 12]

"In a further embodiment, comparative gene transcript frequency analysis is used . . . for the selection of better pharmacologic animal models." [page 14]

"In a further embodiment, comparative gene transcript frequency analysis is used in a clinical setting to <u>give a highly detailed gene transcript profile</u> of a diseased state or condition." [page 14]

"An alternate method of producing a gene transcript image includes the steps of obtaining a mixture of test mRNA and providing a representative array of unique probes whose sequences are complementary to at least some of the test mRNAs. Next, a fixed amount of the test mRNA is added to the arrayed probes. The test mRNA is incubated with the probes for a sufficient time to allow hybrids of the test mRNA and probes to form. The mRNA-probe hybrids are detected and the quantity determined." [page 15]

"[T]his research tool provides a way to get new drugs to the public faster and more economically." [page 36]

"In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a clinical marker." [page 38]

"[T]he gene transcript changes noted in the earlier rat toxicity study are carefully evaluated as clinical markers in the followed patients. Changes in the gene transcript image analyses are evaluated as indicators of toxicity by correlation with clinical signs and symptoms and other laboratory results. . . . The . . . analysis highlights any toxicological changes in the treated patients." [page 39]

U.S. Pat. No. 5,569,588 ("Methods for Drug Screening") ("the '588 patent"), issued October 29, 1996, with a priority date of August 1995, describes an expression profiling platform, the "genome reporter matrix", which is different from nucleic acid microarrays. Additionally describing use of nucleic acid microarrays, the 588 patent makes clear that the utility of comparing multidimensional expression datasets is independent of the methods by which such profiles are obtained. The 588 patent speaks clearly to the usefulness of such expression analyses in drug development and toxicology, particularly pointing out that a gene's failure to change in expression level is a useful result. Thus, with emphasis added,

The invention provides "[m]ethods and compositions for modeling the transcriptional responsiveness of an organism to a candidate drug. . . . [The final step of the method comprises] comparing reporter gene product signals for each cell before and after contacting the cell with the candidate drug to obtain a drug response profile which provides a model of the transcriptional responsiveness of said organism to the candidate drug." [abstract]

"The present invention exploits the recent advances in genome science to provide for the rapid screening of large numbers of compounds against a systemic target comprising substantially all targets in a pathway [or] organism." [col. 1]

"The ensemble of reporting cells comprises as comprehensive a collection of transcription regulatory genetic elements as is conveniently available for the targeted organism so as to most accurately model the systemic transcriptional response. Suitable ensembles generally comprise thousands of individually reporting elements; preferred ensembles are substantially comprehensive, i.e. provide a transcriptional response diversity comparable to that of the target organism. Generally, a substantially comprehensive ensemble requires transcription regulatory genetic elements from at least a majority of the organism's genes, and preferably includes those of all or nearly all of the genes. We term such a substantially comprehensive ensemble a genome reporter matrix." [col. 2]

"Drugs often have side effects that are in part due to the lack of target specificity. . . . [A] genome reporter matrix reveals the spectrum of other genes in the genome also affected by the compound. In considering two different compounds both of which induce the ERG10 reporter, if one compound affects the expression of 5 other reporters and a second compound affects the expression of 50 other reports, the first compound is, a priori, more likely to have fewer side effects." [cols. 2 - 3]

"Furthermore, it is not necessary to know the identity of any of the responding genes." [col. 3]

"[A]ny new compound that induces the same response profile as [a] . . . dominant tubulin mutant would provide a candidate for a taxol-like pharmaceutical." [col. 4]

"The genome reporter matrix offers a simple solution to recognizing new specificities in combinatorial libraries. Specifically, pools of new compounds are tested as mixtures across the matrix. If the pool has any new activity not present in the original lead compound, new genes are affected among the reporters." [col. 4]

" A sufficient number of different recombinant cells are included to provide an ensemble of transcriptional regulatory elements of said organism sufficient to model the transcriptional responsiveness of said organism to a drug. In a preferred embodiment, the matrix is substantially comprehensive for the selected regulatory elements, e.g. essentially all of the gene promoters of the targeted organism are included." [cols. 6 – 7]

"In a preferred embodiment, the basal response profiles are determined. . . . The resultant electrical output signals are stored in a computer memory as genome reporter output signal matrix data structure associating each output signal with the coordinates of the corresponding microtiter plate well and the stimulus or drug. This information is indexed against the matrix to form reference response profiles that are used to determine the response of each reporter to any milieu in which a stimulus may be provided. After establishing a basal response profile for the matrix, each cell is contacted with a candidate drug. The term drug is used loosely

to refer to agents which can provoke a specific cellular response. . . . <u>The drug induces a complex response pattern</u> of repression, silence and induction across the matrix . . . .The response profile reflects the cell's transcriptional adjustments to maintain homeostasis in the presence of the drug. . . . After contacting the cells with the candidate drug, the reporter gene product signals from each of said cells is again measured to determine a stimulated response profile. The basal o[r] background response profile is then compared with . . . the stimulated response profile to identify the cellular response profile to the candidate drug." [cols. 7 - 8]

"<u>In another embodiment of the invention</u>, <u>a matrix [i.e., array] of hybridization probes corresponding to a predetermined population of genes of the selected organism is used</u> to specifically detect changes in gene transcription which result from exposing the selected organism or cells thereof to a candidate drug. In this embodiment, one or more cells derived from the organism is exposed to the candidate drug in vivo or ex vivo under conditions wherein the drug effects a change in gene transcription in the cell to maintain homeostasis. Thereafter, the gene transcripts, primarily mRNA, of the cell or cells is isolated . . . [and] then contacted with an ordered matrix [array] of hybridization probes, each probe being specific for a different one of the transcripts, <u>under conditions where each of the transcripts hybridizes with a corresponding one</u> of the probes to form hybridization pairs. The ordered matrix of probes provides, in aggregate, complements for an ensemble of genes of the organism sufficient to model the transcriptional responsiveness of the organism to a drug. . . . The <u>matrix-wide signal profile of the drug-stimulated cells is then compared with a matrix-wide signal profile of negative control cells to obtain a specific drug response profile."</u> [col. 8]

"The invention also provides means for computer-based qualitative analysis of candidate drugs and unknown compounds. A wide variety of reference response profiles may be generated and used in such analyses." [col. 8]

"<u>Response profiles for an unknown stimulus (e.g. new chemicals, unknown compounds or unknown mixtures) may be analyzed by comparing the new stimulus response profiles with response profiles to known chemical stimuli</u>." [col. 9]

"The response profile of a new chemical stimulus may also be compared to a known genetic response profile for target gene(s)." [col. 9]

The August 11, 1997 press release from the '588 patent's assignee, Acacia Biosciences (now part of Merck) (reference "9h" attached hereto), and the September 15, 1997 news report by Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," *Genetic Engineering News* (reference "9i" attached hereto), attest the commercial value of the methods and technology described and claimed in the '588 patent.

WO 97/13877 ("Measurement of Gene Expression Profiles in Toxicity

Determinations"), published April 17, 1997, describes an expression profiling technology

differing somewhat from the use of cDNA microarrays and differing from the genome reporter

matrix of the '588 patent; but the use of the data is analogous. As per its title, the reference

describes use of expression profiling in toxicity determinations. In particular, and with emphasis

added:

> "[T]he invention relates to a method for detecting and monitoring changes
> in gene expression patterns in in vitro and in vivo systems for determining the
> toxicity of drug candidates." [Field of the invention]

> "An object of the invention is to provide <u>a new approach to toxicity
> assessment based on an examination of gene expression patterns, or profiles</u>, in in
> vitro or in vivo test systems." [page 3]

> "Another object of the invention is to provide a rapid and reliable method
> for correlating gene expression with short term and long term toxicity in test
> animals." [page 3]

> "The invention achieves these and other objects by providing a method for
> massively parallel signature sequencing of genes expressed in one or more selected
> tissues of an organism exposed to a test compound. An important feature of the
> invention is the application of novel . . . methodologies that permit the formation of
> gene expression profiles for selected tissues . . . . Such <u>profiles may be compared
> with those from tissues of control organisms at single or multiple time points <u>to</u>
> <u>identify expression patterns predictive of toxicity</u>." [page 3]

> "As used herein, the terms 'gene expression profile,' and 'gene expression
> pattern' which is used equivalently, means a frequency distribution of sequences of
> portions of cDNA molecules sampled from a population of tag-cDNA conjugates. .
> .. Preferably, the total number of sequences determined is at least 1000; <u>more
> preferably, the total number of sequences determined in a gene expression profile is
> at least ten thousand</u>." [page 7]

> "The invention provides a method for determining the toxicity of a
> compound by analyzing changes in the gene expression profiles in selected tissues
> of test organisms exposed to the compound. . . . . Gene expression profiles derived
> from test organisms are compared to gene expression profiles derived from control
> organisms. . . . " [page 7]

Therefore, the potential benefit to the public, in terms of lives saved and reduced health

care costs, are enormous. Evidence of the benefits of this information include:

- In 1999, CV Therapeutics, an Incyte collaborator, was able to use Incyte gene expression technology, information about the structure of a known transporter gene, and chromosomal mapping location, to identify the key gene associated with Tangiers disease. This discovery took place over a matter of only a few weeks, due to the power of these new genomics technologies. The discovery received an award from the American Heart Association as one of the top 10 discoveries associated with heart disease research in 1999.

- In an April 9, 2000, article published by the Bloomberg news service, an Incyte customer stated that it had reduced the time associated with target discovery and validation from 36 months to 18 months, through use of Incyte's genomic information database. Other Incyte customers have privately reported similar experiences. The implications of this significant saving of time and expense for the number of drugs that may be developed and their cost are obvious.

- In a February 10, 2000, article in the *Wall Street Journal*, one Incyte customer stated that over 50 percent of the drug targets in its current pipeline were derived from the Incyte database. Other Incyte customers have privately reported similar experiences. By doubling the number of targets available to pharmaceutical researchers, Incyte genomic information has demonstrably accelerated the development of new drugs.

Because the Patent Examiner failed to address or consider the "well-established" utilities for the claimed invention in toxicology testing, drug development, and the diagnosis of disease, the Examiner's rejections should be overturned regardless of their merit.

C.   The uncontested fact that the claimed polynucleotide encodes a protein in the GPCR family also demonstrates utility

In addition to having substantial, specific and credible utilities in numerous gene expression monitoring applications, it is undisputed that the claimed polynucleotide encodes for a protein having the sequence shown as SEQ ID NO:1 in the patent application. Appellants have demonstrated that SEQ ID NO:1 is a member of the GPCR family, and that the GPCR family of proteins includes glutamate GPCRs that function in neurotransmission, and play a role in certain neurological diorders.

The Patent Examiner does not dispute any of the facts set forth in the previous paragraph. Neither does the Patent Examiner dispute that, if a polynucleotide encodes for a protein that has a substantial, specific and credible utility, then it follows that the polynucleotide also has a

substantial, specific and credible utility.

The Examiner must accept the applicant's demonstration that the polypeptide encoded by the claimed invention is a member of the GPCR family and that utility is proven by a reasonable probability unless the Examiner can demonstrate through evidence or sound scientific reasoning that a person of ordinary skill in the art would doubt utility. *See In re Langer*, 503 F.2d 1380, 1391-92, 183 USPQ 288 (CCPA 1974). The Examiner has not provided sufficient evidence or sound scientific reasoning to the contrary.

Nor has the Examiner provided any evidence that any member of the GPCR family, let alone a substantial number of those members, is not useful. In such circumstances, the only reasonable inference is that the polypeptide encoded by the claimed invention must be, like the other members of the GPCR family, useful.

D. Objective evidence corroborates the utilities of the claimed invention

There is, in fact, no restriction on the kinds of evidence a Patent Examiner may consider in determining whether a "real-world" utility exists. "Real-world" evidence, such as evidence showing actual use or commercial success of the invention, can demonstrate conclusive proof of utility. *Raytheon v. Roper*, 220 USPQ2d 592 (Fed. Cir. 1983); *Nestle v. Eugene*, 55 F.2d 854, 856, 12 USPQ 335 (6th Cir. 1932). Indeed, proof that the invention is made, used or sold by any person or entity other than the patentee is conclusive proof of utility. *United States Steel Corp. v. Phillips Petroleum Co.*, 865 F.2d 1247, 1252, 9 USPQ2d 1461 (Fed. Cir. 1989).

Over the past several years, a vibrant market has developed for databases containing the sequences of all expressed genes (along with the polypeptide translations of those genes), in particular genes having medical and pharmaceutical significance such as the instant sequence. (Note that the value in these databases is enhanced by their completeness, but each sequence in them is independently valuable.) The databases sold by Appellants' assignee, Incyte, include exactly the kinds of information made possible by the claimed invention, such as tissue and disease associations. Incyte sells its database containing the claimed sequence and millions of other sequences throughout the scientific community, including to pharmaceutical companies who use the information to develop new pharmaceuticals.

Both Incyte's customers and the scientific community have acknowledged that Incyte's

databases have proven to be valuable in, for example, the identification and development of drug candidates. Page et al., in discussing the identification and assignment of candidate drug targets, state that "rapid identification and assignment of candidate targets and markers represents a huge challenge ... [t]he process of annotation is similarly aided by the quantity and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals)" Page, M.J. et al., "Proteomics: a major new technology for the drug discovery process," Drug Discov. Today 4:55-62 (1999) (Reference No. 6), see page 58, col. 2). As Incyte adds information to its databases, including the information that can be generated only as a result of Incyte's invention of the claimed polynucleotide and its use of that polynucleotide on cDNA microarrays, the databases become even more powerful tools. Thus the claimed invention adds more than incremental benefit to the drug discovery and development process.

III. The Patent Examiner's rejections are without merit

Rather than responding to the evidence demonstrating utility, the Examiner attempts to dismiss it altogether by arguing that the disclosed and well-established utilities for the claimed polynucleotide are not "specific, substantial, and credible" utilities. (Final Office Action at page 3). The Examiner is incorrect both as a matter of law and as a matter of fact.

A. The precise biological role or function of an expressed polynucleotide is not required to demonstrate utility

The Patent Examiner's primary rejection of the claimed invention is based on the ground that, without information as to the precise "biological role" of the claimed invention, the claimed invention's utility is not sufficiently specific. According to the Examiner, it is not enough that a person of ordinary skill in the art could use and, in fact, would want to use the claimed invention either by itself or in a cDNA microarray to monitor the expression of genes for such applications as the evaluation of a drug's efficacy and toxicity. The Examiner would require, in addition, that the applicant provide a specific and substantial interpretation of the results generated in any given expression analysis.

It may be that specific and substantial interpretations and detailed information on

biological function are necessary to satisfy the requirements for publication in some technical journals, but they are not necessary to satisfy the requirements for obtaining a United States patent. The relevant question is not, as the Examiner would have it, whether it is known how or why the invention works, *In re Cortwright*, 165 F.3d 1353, 1359 (Fed. Cir. 1999), but rather whether the invention provides an "identifiable benefit" in presently available form. *Juicy Whip Inc.* v. *Orange Bang Inc.*, 185 F.3d 1364, 1366 (Fed. Cir. 1999). If the benefit exists, and there is a substantial likelihood the invention provides the benefit, it is useful. There can be no doubt, particularly in view of the First Bedilion Declaration (at, *e.g.*, ¶¶ 10 and 15), that the present invention meets this test.

The threshold for determining whether an invention produces an identifiable benefit is low. *Juicy Whip*, 185 F.3d at 1366. Only those utilities that are so nebulous that a person of ordinary skill in the art would not know how to achieve an identifiable benefit and, at least according to the PTO guidelines, so-called "throwaway" utilities that are not directed to a person of ordinary skill in the art at all, do not meet the statutory requirement of utility. Utility Examination Guidelines, 66 Fed. Reg. 1092 (Jan. 5, 2001).

Knowledge of the biological function or role of a biological molecule has never been required to show real-world benefit. In its most recent explanation of its own utility guidelines, the PTO acknowledged as much (66 F.R. at 1095):

> [T]he utility of a claimed DNA does not necessarily depend on the function of the encoded gene product. A claimed DNA may have specific and substantial utility because, *e.g.*, it hybridizes near a disease-associated gene or it has gene-regulating activity.

By implicitly requiring knowledge of biological function for any claimed nucleic acid, the Examiner has, contrary to law, elevated what is at most an evidentiary factor into an absolute requirement of utility. Rather than looking to the biological role or function of the claimed invention, the Examiner should have looked first to the benefits it is alleged to provide.

B.   Membership in a class of useful products can be proof of utility

Despite the uncontradicted evidence that the claimed polynucleotide encodes a polypeptide in the GPCR family, the Examiner refused to impute the utility of the members of the GPCR family to SEQ ID NO:1. In the Final Office Action, the Patent Examiner takes the

position that, unless Appellants can identify which particular biological function within the class of GPCRs is possessed by SEQ ID NO:1, utility cannot be imputed. See Final Office Action, page 4. To demonstrate utility by membership in the class of GPCRs, the Examiner would require that all GPCRs possess a "common" utility.

There is no such requirement in the law. In order to demonstrate utility by membership in a class, the law requires only that the class not contain a substantial number of useless members. So long as the class does not contain a substantial number of useless members, there is sufficient likelihood that the claimed invention will have utility, and a rejection under 35 U.S.C. § 101 is improper. That is true regardless of how the claimed invention ultimately is used and whether or not the members of the class possess one utility or many. See *Brenner v. Manson*, 383 U.S. 519, 532 (1966); *Application of Kirk*, 376 F.2d 936, 943 (CCPA 1967).

Membership in a "general" class is insufficient to demonstrate utility only if the class contains a sufficient number of useless members such that a person of ordinary skill in the art could not impute utility by a substantial likelihood. There would be, in that case, a substantial likelihood that the claimed invention is one of the useless members of the class. In the few cases in which class membership did not prove utility by substantial likelihood, the classes did in fact include predominately useless members. *E.g.*, *Brenner* (man-made steroids); *Kirk* (same); *Natta* (man-made polyethylene polymers).

The Examiner addresses GPCRs as if the general class in which it is included is not the GPCR family, but rather all polynucleotides or all polypeptides, including the vast majority of useless theoretical molecules not occurring in nature, and thus not pre-selected by nature to be useful. While these "general classes" may contain a substantial number of useless members, the GPCR family does not. The GPCR family is sufficiently specific to rule out any reasonable possibility that SEQ ID NO:1 would not also be useful like the other members of the family.

Because the Examiner has not presented any evidence that the GPCR class of signaling molecules has any, let alone a substantial number, of useless members, the Examiner must conclude that there is a "substantial likelihood" that the SEQ ID NO:1 encoded by the claimed polynucleotide is useful. It follows that the claimed polynucleotide also is useful.

C.   **Because the uses of the claimed polynucleotide in toxicology testing, drug discovery, and disease diagnosis are practical uses beyond mere study of the**

invention itself, the claimed invention has substantial utility

The Examiner's rejection of the claims at issue as not having a "substantial" use is tantamount to a rejection based on an allegation that the only use of the claimed invention is as a tool for further research. Because the PTO's rejection assumes a substantial overstatement of the law, and is incorrect in fact, it must be overturned.

There is no authority for the proposition that use as a tool for research is not a substantial utility. Indeed, the Patent Office has recognized that just because an invention is used in a research setting does not mean that it lacks utility (Section § 2107.01 of the Manual of Patent Examining Procedure, 8th Edition, August 2001, under the heading I. Specific and Substantial Requirements, Research Tools):

> Many research tools such as gas chromatographs, screening assays, and nucleotide sequencing techniques have a clear, specific and unquestionable utility (e.g., they are useful in analyzing compounds). An assessment that focuses on whether an invention is useful only in a research setting thus does not address whether the specific invention is in fact "useful" in a patent sense. Instead, Office personnel must distinguish between inventions that have a specifically identified utility and inventions whose specific utility requires further research to identify or reasonably confirm.

The Patent Office's actual practice has been, at least until the present, consistent with that approach. It has routinely issued patents for inventions whose only use is to facilitate research, such as DNA ligases. These are acknowledged by the PTO's Training Materials themselves to be useful, as well as DNA sequences used, for example, as markers.

Only a limited subset of research uses are not "substantial" utilities: those in which the only known use for the claimed invention is to be an **object** of further study, thus merely inviting further research. This follows from *Brenner*, in which the U.S. Supreme Court held that a process for making a compound does not confer a substantial benefit where the only known use of the compound was to be the object of further research to determine its use. *Id.* at 535. Similarly, in *Kirk*, the Court held that a compound would not confer substantial benefit on the public merely because it might be used to synthesize some other, unknown compound that would confer substantial benefit. *Kirk*, 376 F.2d at 940, 945 ("What appellants are really saying to those in the art is take these steroids, experiment, and find what use they do have as medicines."). Nowhere do those cases state or imply, however, that a material cannot be patentable if it has some other beneficial use in research.

D. The Patent Examiner failed to demonstrate that a person of ordinary skill in the art would reasonably doubt the utility of the claimed invention

The Examiner alleges that applicants asserted use of the claimed polynucleotide in the detection and diagnosis of cancer, in particular, thyroid cancer, is based on a correlation with thyroid cancer in on a single library representing follicular carcinoma of the thyroid. See specification, at page 35. Applicants reiterate that the asserted utility for the polynucleotide encoding SEQ ID NO:1 in the detection and diagnosis of follicular carcinoma of the thyroid, based on a significant (4-fold) differential expression in that disease condition, is both specific, substantial, and credible. The Examiners' allegation that the asserted utility is not credible because it is based on expression of the transcript in only one library ignores the fact that a number of thyroid libraries were examined representing both normal and diseased thyroid, and that only libraries associated with thyroid cancer were found to express the gene. In particular, the gene was most highly expressed in a thyroid follicular carcinoma tumor library (THYRTUP02), but was also expressed in a library associated with follicular adenoma (THYRNOT03), a precancerous condition to follicular carcinoma. Such evidence provides more than a "substantial likelihood" that the polynucleotide may be used in the detection and diagnosis of the disease. Further, the evidence provided from the Northern analysis for SEQ ID NO:7 supports applicants assertion or the use of the claimed polynucleotide in cancer as disclosed in the Bandman '513 priority application at pages 29-30. The Examiners' reliance on references such as the NCI Guidelines for Marker Development to support her position is merely an attempt to raise the standard for utility to one of near certainty. However, the standard applicable in this case is not proof to certainty, but rather proof to reasonable probability. *Brenner*, 383 U.S. at 532.

Applicants' Showing of Facts Overcomes The Examiner's Concern That Applicants' Invention Lacks "Specific Utility"

The Examiner alleges that the claimed invention is not supported by either a specific and substantial asserted utility or a well established utility. (Final Office Action, page 3.)

Appellants' submission of additional facts overcomes this concern. Those facts demonstrate that, far from applying *regardless* of the specific properties of the claimed

invention, the utility of Appellants' claimed polynucleotides as gene-specific probes *depends upon* specific properties of the polynucleotides, that is, their nucleic acid sequences.

"[E]ach probe on . . . [a "high density spotted microarray[]"], with careful design and sufficient length, and with sufficiently stringent hybridization and wash conditions, *binds specifically* and with minimal cross-hybridization, to the probe's cognate transcript" [1], "[e]ach gene included as a probe on a microarray provides *a signal that is specific to the cognate transcript*, at least to a first approximation." [2] Accordingly, "each additional probe makes an additional transcript newly detectable by the microarray, increasing the detection range, and thus versatility, of this analytical device for gene expression profiling" [3], equally, "[e]ach new gene-specific probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device." [4]

Although not required for present purposes, it would be appropriate to state on the record here that the specificity of nucleic acid hybridization was well-established far earlier than the development of high density spotted microarrays in 1995, and indeed is the well-established underpinning of many, perhaps most, molecular biological techniques developed over the past 30 - 40 years.

IV. By requiring the patent applicant to assert a particular or unique utility, the Patent Examination Utility Guidelines and Training Materials applied by the Patent Examiner misstate the law

There is an additional, independent reason to overturn the rejections: to the extent the rejections are based on Revised Interim Utility Examination Guidelines (64 FR 71427, December 21, 1999), the final Utility Examination Guidelines (66 FR 1092, January 5, 2001) and/or the Revised Interim Utility Guidelines Training Materials (USPTO Website www.uspto.gov, March 1, 2000), the Guidelines and Training Materials are themselves inconsistent with the law.

---

[1] Declaration of Dr. John C. Rockett, ¶ 10(i), emphasis added.

[2] Declaration of Dr. Vishwanath R. Iyer, ¶ 7 (emphasis added). See the footnote at ¶ 7 for a slightly more "nuanced" view.

[3] Declaration of Dr. John C. Rockett, ¶ 10(ii).

[4] Declaration of Dr. Vishwanath R. Iyer, ¶ 7.

The Training Materials, which direct the Examiners regarding how to apply the Utility Guidelines, address the issue of specificity with reference to two kinds of asserted utilities: "specific" utilities which meet the statutory requirements, and "general" utilities which do not. The Training Materials define a "specific utility" as follows:

> A [specific utility] is *specific* to the subject matter claimed. This contrasts to *general* utility that would be applicable to the broad class of invention. For example, a claim to a polynucleotide whose use is disclosed simply as "gene probe" or "chromosome marker" would not be considered to be specific in the absence of a disclosure of a specific DNA target. Similarly, a general statement of diagnostic utility, such as diagnosing an unspecified disease, would ordinarily be insufficient absent a disclosure of what condition can be diagnosed.

The Training Materials distinguish between "specific" and "general" utilities by assessing whether the asserted utility is sufficiently "particular," *i.e.*, unique (Training Materials at page 52) as compared to the "broad class of invention." (In this regard, the Training Materials appear to parallel the view set forth in Stephen G. Kunin, <u>Written Description Guidelines and Utility Guidelines</u>, 82 J.P.T.O.S. 77, 97 (Feb. 2000) ("With regard to the issue of specific utility the question to ask is whether or not a utility set forth in the specification is *particular* to the claimed invention.")).

Such "unique" or "particular" utilities never have been required by the law. To meet the utility requirement, the invention need only be "practically useful," *Natta*, 480 F.2d 1 at 1397, and confer a "specific benefit" on the public. *Brenner*, 383 U.S. at 534. Thus, incredible "throw-away" utilities, such as trying to "patent a transgenic mouse by saying it makes great snake food," do not meet this standard. Karen Hall, <u>Genomic Warfare</u>, The American Lawyer 68 (June 2000) (quoting John Doll, Chief of the Biotech Section of USPTO).

This does not preclude, however, a general utility, contrary to the statement in the Training Materials where "specific utility" is defined (page 5). Practical real-world uses are not limited to uses that are unique to an invention. The law requires that the practical utility be "definite," not particular. *Montedison*, 664 F.2d at 375. Appellants are not aware of any court that has rejected an assertion of utility on the grounds that it is not "particular" or "unique" to the specific invention. Where courts have found utility to be too "general," it has been in those cases in which the asserted utility in the patent disclosure was not a practical use that conferred a

specific benefit. That is, a person of ordinary skill in the art would have been left to guess as to how to benefit at all from the invention. In *Kirk*, for example, the CCPA held the assertion that a man-made steroid had "useful biological activity" was insufficient where there was no information in the specification as to how that biological activity could be practically used. *Kirk*, 376 F.2d at 941.

The fact that an invention can have a particular use does not provide a basis for requiring a particular use. See *Brana, supra* (disclosure describing a claimed antitumor compound as being homologous to an antitumor compound having activity against a "particular" type of cancer was determined to satisfy the specificity requirement). "Particularity" is not and never has been the *sine qua non* of utility; it is, at most, one of many factors to be considered.

As described *supra*, broad classes of inventions can satisfy the utility requirement so long as a person of ordinary skill in the art would understand how to achieve a practical benefit from knowledge of the class. Only classes that encompass a significant portion of nonuseful members would fail to meet the utility requirement. Montedison, 664 F.2d at 374-75.

The Training Materials fail to distinguish between broad classes that convey information of practical utility and those that do not, lumping all of them into the latter, unpatentable category of "general" utilities. As a result, the Training Materials paint with too broad a brush. Rigorously applied, they would render unpatentable whole categories of inventions that heretofore have been considered to be patentable and that have indisputably benefitted the public, including the claimed invention. See *supra* § II.B. Thus the Training Materials cannot be applied consistently with the law.

V.   **To the extent the rejection of the claimed invention under 35 U.S.C. § 112, first paragraph, is based on the improper rejection for lack of utility under 35 U.S.C. § 101, it must be reversed.**

The rejection set forth in the Office Action is based on the assertions discussed above, i.e., that the claimed invention lacks patentable utility. To the extent that the rejection under 35 U.S.C. § 112, first paragraph, is based on the improper allegation of lack of patentable utility under 35 U.S.C. § 101, it fails for the same reasons.

CONCLUSION

Appellants respectfully submit that rejections for lack of utility based, *inter alia*, on an allegation of "lack of specificity," as set forth in the Office Action and as justified in the Revised Interim and final Utility Guidelines and Training Materials, are not supported in the law. Neither are they scientifically correct, nor supported by any evidence or sound scientific reasoning. These rejections are alleged to be founded on facts in court cases such as *Brenner* and *Kirk*, yet those facts are clearly distinguishable from the facts of the instant application, and indeed most if not all nucleotide and protein sequence applications. Nevertheless, the PTO is attempting to mold the facts and holdings of these prior cases, "like a nose of wax," [5] to target rejections of claims to polypeptide and polynucleotide sequences, where biological activity information has not been proven by laboratory experimentation, and they have done so by ignoring perfectly acceptable utilities fully disclosed in the specifications as well as well-established utilities known to those of skill in the art. As is disclosed in the specification, and even more clearly, as one of ordinary skill in the art would understand, the claimed invention has well-established, specific, substantial and credible utilities. The rejections are, therefore, improper and should be reversed.

Moreover, to the extent the above rejections were based on the Revised Interim and final Examination Guidelines and Training Materials, those portions of the Guidelines and Training Materials that form the basis for the rejections should be determined to be inconsistent with the law.

Claims 1-6 stand rejected under 35 U.S.C. § 112, first paragraph, as containing subject matter which is not described in the specification in such a was as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention. The rejection alleges in particular, that:

- while the specification describes a polypeptide sequence consisting of SEQ ID NO:1, the claims encompass polypeptides comprising fragments and homologues that vary

---

[5] "The concept of patentable subject matter under §101 is not 'like a nose of wax which may be turned and twisted in any direction * * *.' *White v. Dunbar*, 119 U.S. 47, 51." *(Parker v. Flook*, 198 USPQ 193 (US SupCt 1978))

substantially in length and also in amino acid composition. The instant disclosure of a single polypeptide, that of SEQ ID NO:1, does not support the scope of the claimed genus, which encompasses a substantial variety of subgenera. See *Reagents of the University of California v Eli Lilly* with respect to the premise that "A description of a genus of cDNAs may be achieved by means of a recitation of a representative number of cDNAs, defined by nucleotide sequence, falling within the scope of the genus, or a recitation of structural features common to the genus, which features constitute a substantial portion of the genus". The Examiner then cited various references alleging to support the unpredictability of protein function based on sequence homology. See, in particular, Vukicevic et al.; Tischer et al.; and Kopchick et al. The Examiner concluded by saying that given the unpredictability of homology comparisons, and the fact that the specification fails to provide objective evidence that the additional sequences are indeed species of the claimed genus, it cannot be established that a representative number of species have been disclosed by the claims. Further, the Examiner stated, no activity is set forth for the additional sequences.

The recited fragments and variants of SEQ ID NO:1 and SEQ ID NO:2 are sufficiently described in chemical and structural terms that the skilled artisan would recognize applicant's possession of them at the time the application was filed

With respect to fragments of SEQ ID NO:1, as recited in claim 1, applicants submit that the recited fragments are disclosed in the specification and claims in terms of their specific amino acid sequences and therefore clearly meet the requirements for written description under 35 U.S.C. § 112, first paragraph..

The claimed "homologues" of SEQ ID NO:1 referred to by the Examiner presumably relate to variants of SEQ ID NO:1 and SEQ ID NO:7, as recited in claims 1 and 2, respectively. Applicants submit that the polypeptides and polynucleotides of the invention, including the recited variants, are adequately described in accordance with 35 U.S.C. § 112, first paragraph, and supported by relevant case law, some of which is referred to by the Examiner.

The requirements necessary to fulfill the written description requirement of 35 U.S.C. 112, first paragraph, are well established by case law.

. . . the applicant must also convey with reasonable clarity to those skilled in the art that, as of the filing date sought, he or she was in possession *of the invention*. The invention is, for purposes of the "written description" inquiry, *whatever is now claimed*. *Vas-Cath, Inc. v. Mahurkar*, 19 USPQ2d 1111, 1117 (Fed. Cir. 1991)

Attention is also drawn to the Patent and Trademark Office's own "Guidelines for Examination of Patent Applications Under the 35 U.S.C. Sec. 112, para. 1", published January 5, 2001, which provide that :

An applicant may also show that an invention is complete by disclosure of sufficiently detailed, relevant identifying characteristics which provide evidence that applicant was in possession of the claimed invention, i.e., complete or partial structure, other physical and/or chemical properties, functional characteristics when coupled with a known or disclosed correlation between function and structure, or some combination of such characteristics. What is conventional or well known to one of ordinary skill in the art need not be disclosed in detail. If a skilled artisan would have understood the inventor to be in possession of the claimed invention at the time of filing, even if every nuance of the claims is not explicitly described in the specification, then the adequate description requirement is met.

Thus, the written description standard is fulfilled by both what is specifically disclosed and what is conventional or well known to one skilled in the art.

SEQ ID NO:1 and SEQ ID NO:7 are specifically disclosed in the priority application Serial No. 09/156,513 (see, for example, page 2, lines 34-37 and page 3, lines 13-14). Variants of SEQ ID NO:1 and SEQ ID NO:7 are described, for example, at page 2, line 38 through page 3, line 2. In particular, the preferred, more preferred, and most preferred variants (80%, 90%, and 95% amino acid sequence similarity to SEQ ID NO:1) are described, for example, at page 12, lines 13-16 of priority application Serial No. 09/156,513. Incyte clones in which the nucleic acids encoding the human HGPRP-1 (SEQ ID NO:1) were first identified and libraries from which those clones were isolated are described, for example, at page 11, lines 24-30 and Table 1 of the priority application. Chemical and structural features of SEQ ID NO:1 are described, for example, on page 11, lines 31-35 and Table 2 of the priority application. Given SEQ ID NO:1, one of ordinary skill in the art would recognize naturally-occurring variants of SEQ ID NO:1

having at least 90% sequence identity to SEQ ID NO:1. Accordingly, the Specification provides an adequate written description of the recited polypeptide sequences.

A. The Specification provides an adequate written description of the claimed "variants" of SEQ ID NO:1.

The Office Action has further asserted that the claims are not supported by an adequate written description because:

> Claims 1-6 contain "subject matter which is not described in the specification in such a was as to reasonably convey to one skilled in the relevant art that the inventor(s), at the time the application was filed, had possession of the claimed invention".

(page 8 of the Final Office Action)

Such a position is believed to present a misapplication of the law.

1. The present claims specifically define the claimed genus through the recitation of chemical structure

Court cases in which "DNA claims" have been at issue (which are hence relevant to claims to proteins encoded by the DNA and antibodies which specifically bind to the proteins) commonly emphasize that the recitation of structural features or chemical or physical properties are important factors to consider in a written description analysis of such claims. For example, in *Fiers v. Revel*, 25 USPQ2d 1601, 1606 (Fed. Cir. 1993), the court stated that:

> If a conception of a DNA requires a precise definition, such as by structure, formula, chemical name or physical properties, as we have held, then a description also requires that degree of specificity.

In a number of instances in which claims to DNA have been found invalid, the courts have noted that the claims attempted to define the claimed DNA in terms of functional characteristics without any reference to structural features. As set forth by the court in *University of California v. Eli Lilly and Co.*, 43 USPQ2d 1398, 1406 (Fed. Cir. 1997):

> In claims to genetic material, however, a generic statement such as "vertebrate insulin cDNA" or "mammalian insulin cDNA," without more, is not an adequate written description of the genus because it does not distinguish the claimed genus from others, except by function.

Thus, the mere recitation of functional characteristics of a DNA, without the definition of structural features, has been a common basis by which courts have found invalid claims to DNA. For example, in *Lilly*, 43 USPQ2d at 1407, the court found invalid for violation of the written description requirement the following claim of U.S. Patent No. 4,652,525:

> 1. A recombinant plasmid replicable in procaryotic host containing within its nucleotide sequence a subsequence having the structure of the reverse transcript of an mRNA of a vertebrate, which mRNA encodes insulin.

In *Fiers*, 25 USPQ2d at 1603, the parties were in an interference involving the following count:

> A DNA which consists essentially of a DNA which codes for a human fibroblast interferon-beta polypeptide.

Party Revel in the *Fiers* case argued that its foreign priority application contained an adequate written description of the DNA of the count because that application mentioned a potential method for isolating the DNA. The Revel priority application, however, did not have a description of any particular DNA structure corresponding to the DNA of the count. The court therefore found that the Revel priority application lacked an adequate written description of the subject matter of the count.

Thus, in *Lilly* and *Fiers*, nucleic acids were defined on the basis of functional characteristics and were found not to comply with the written description requirement of 35 U.S.C. §112; *i.e.*, "an mRNA of a vertebrate, which mRNA encodes insulin" in *Lilly*, and "DNA which codes for a human fibroblast interferon-beta polypeptide" in *Fiers*. In contrast to the situation in *Lilly* and *Fiers*, the claims at issue in the present application define polynucleotides and polypeptides in terms of chemical structure, rather than functional characteristics. For example, the "variant language" of independent claim 1 recites chemical structure to define the claimed genus:

> 1. An isolated cDNA comprising a nucleic acid encoding an amino acid sequence selected from:....c) a variant of SEQ ID NO:1 having at least 90% amino acid sequence identity to SEQ ID NO:1...

From the above it should be apparent that the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the

present claims is defined in terms of the chemical structure of SEQ ID NO:1. In the present case, there is no reliance merely on a description of functional characteristics of the polynucleotides or polypeptides recited by the claims. In fact, there is no recitation of functional characteristics. Moreover, if such functional recitations were included, it would add to the structural characterization of the recited polynucleotides or polypeptides or. The polynucleotides or polypeptides defined in the claims of the present application recite structural features, and cases such as *Lilly* and *Fiers* stress that the recitation of structure is an important factor to consider in a written description analysis of claims of this type. By failing to base its written description inquiry "on whatever is now claimed," the Office Action failed to provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in *Lilly* and *Fiers*

2.   The present claims do not define a genus which is "highly variant"

Furthermore, the claims at issue do not describe a genus which could be characterized as highly variant, i.e., "encompassing a substantial variety of subgenera" (Final Office Action, page 8). Available evidence illustrates that the claimed genus is of narrow scope.

In support of this assertion, the Examiner's attention is directed to the reference by Brenner et al. ("Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships," Proc. Natl. Acad. Sci. USA (1998) 95:6073-6078; cited at page 29 of the instant application). Through exhaustive analysis of a data set of proteins with known structural and functional relationships and with <90% overall sequence identity, Brenner et al. have determined that 30% identity is a reliable threshold for establishing evolutionary homology between two sequences aligned over at least 150 residues. (Brenner et al., pages 6073 and 6076.) Furthermore, local identity is particularly important in this case for assessing the significance of the alignments, as Brenner et al. further report that ≥40% identity over at least 70 residues is reliable in signifying homology between proteins. (Brenner et al., page 6076.)

The present application is directed, *inter alia*, to GPCR proteins, in particular, metabotropic glutamate GPCR proteins related to the amino acid sequence of SEQ ID NO:1. In accordance with Brenner et al, naturally occurring molecules may exist which could be characterized as metabotropic glutamate GPCR proteins and which have as little as 40% identity over at least 70 residues to SEQ ID NO:1. The "variant language" of the present claims recites, for example, polynucleotides encoding "an amino acid sequence having at least 90% amino acid sequence identity SEQ ID NO:1" (note that SEQ ID NO:1 has 441 amino acid residues). This

variation is far less than that of all potential metabotropic glutamate GPCR proteins related to SEQ ID NO:1, i.e., those metabotropic glutamate GPCR proteins having as little as 40% identity over at least 70 residues to SEQ ID NO:1.

3.   **The state of the art at the time of the present invention is further advanced than at the time of the *Lilly* and *Fiers* applications**

In the *Lilly* case, claims of U.S. Patent No. 4,652,525 were found invalid for failing to comply with the written description requirement of 35 U.S.C. §112. The '525 patent claimed the benefit of priority of two applications, Application Serial No. 801,343 filed May 27, 1977, and Application Serial No. 805,023 filed June 9, 1977. In the *Fiers* case, party Revel claimed the benefit of priority of an Israeli application filed on November 21, 1979. Thus, the written description inquiry in those case was based on the state of the art at essentially at the "dark ages" of recombinant DNA technology.

The present application has a priority date of September 17, 1998. Much has happened in the development of recombinant DNA technology in the 20 or more years from the time of filing of the applications involved in *Lilly* and *Fiers* and the present application. For example, the technique of polymerase chain reaction (PCR) was invented. Highly efficient cloning and DNA sequencing technology has been developed. Large databases of protein and nucleotide sequences have been compiled. Much of the raw material of the human and other genomes has been sequenced. With these remarkable advances one of skill in the art would recognize that, given the sequence information of SEQ ID NO:1 and SEQ ID NO:7, and the additional extensive detail provided by the subject application, the present inventors were in possession of the claimed polynucleotide variants at the time of filing of this application.

4.   **Summary**

The Office Action failed to base its written description inquiry "on whatever is now claimed." Consequently, the Action did not provide an appropriate analysis of the present claims and how they differ from those found not to satisfy the written description requirement in cases such as *Lilly* and *Fiers*. In particular, the claims of the subject application are fundamentally different from those found invalid in *Lilly* and *Fiers*. The subject matter of the present claims is defined in terms of the chemical structure of SEQ ID NO:1 or SEQ ID NO:7. The courts have stressed that structural features are important factors to consider in a written description analysis of claims to nucleic acids and proteins. In addition, the genus of polynucleotides or polypeptides defined by the present claims is adequately described, as evidenced by Brenner et al and

consideration of the claims of the '740 patent involved in *Lilly*. Furthermore, there have been remarkable advances in the state of the art since the *Lilly* and *Fiers* cases, and these advances were given no consideration whatsoever in the position set forth by the Office Action.

Claims 1 and 3-6 stand rejected under 35 U.S.C. § 102(b) as anticipated by Valenzuela et al. (WO 99/55271, November 4, 199) and, alternatively under 35 U.S.C. § 102(e) as anticipated by Moore et al. (U.S. Published Application 2003005536, effective filing date June 17, 1999). The rejection alleges in particular, that:

- Valenzuela disclose a nucleic acid molecule (SEQ ID NO:43, claim 52) that encodes a protein (SEQ ID NO:45, claim 53) that is 100% identical to the polypeptide of SEQ ID NO:7 of the instant application, thus anticipating the claims. Valeuzuela et al. also teach vectors, host cells, a method of producing protein, and labeled cDNA.

- Moore et al. disclose a nucleic acid moecule (SEQ ID NO:22) that encodes a protein (SEQ ID NO:146) that is 100% identical to the polypeptide of SEQ ID NO:1 from amino acids 1-384 of the instant application, and therefore discloses an isolated cDNA encoding a fragment of SEQ ID NO:1 from I51-V72, G88-V109, C116-A145, I156-L175, M207-P229, or G242-T264 of SEQ ID NO:1, as recited in claim 1. Moore et al also teach vectors, host cells, and a method of making a protein, therefore anticipating claims 3-6 as well.

- Because the instant application does not meet the requirements of 35 U.S.C. § 112, first paragraph, for the reasons given above, and it is a continuation of application Serial No. 09/516,513, the prior application does not meet these requirements and therefore is unavailable under 35 U.S.C. § 120. Under these circumstances, Valenzuela et al. and Moore et al. anticipate the claimed invention.

The now claimed invention, at least as recited in claims 1 and 3-6, is supported by both a specific and substantial asserted utility and a well established utility that is disclosed and enabled in priority application Serial No. 09/516,513

Applicants submit that, for the reasons cited above in response to the rejection of claims under 35 U.S.C. §§ 101/112, the specification supports a specific and substantial asserted utility, as well as a well established utility for the claimed invention that is similarly disclosed in the priority application Serial No. 09/516,513 in accordance with 35 U.S.C. § 120, therefore providing an effective filing date for the instant application of September 17, 1998.

Due to the urgency of this matter and its economic and public health implications, an expedited review of this appeal is earnestly solicited.

If the USPTO determines that any additional fees are due, the Commissioner is hereby authorized to charge Deposit Account No. **09-0108.**

This brief is enclosed in triplicate.

Respectfully submitted,

INCYTE CORPORATION

Date: _____January 7, 2004_____

David G. Streeter, Ph.D.
Reg. No. 43,168
Direct Dial Telephone: (650) 845-5741

Customer No.: 27904
3160 Porter Drive
Palo Alto, California 94304
Phone: (650) 855-0555
Fax: (650) 849-8886

Attachments:

1)   Lashkari et al., <u>Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR</u>, Proc. Nat. Acad. Sci. 94:8945-8947

2)   Emile F. Nuwaysir et al., <u>Microarrays and toxicology: The advent of toxicogenomics</u>, Molecular Carcinogenesis 24:153-159 (1999);

3)   Sandra Steiner and N. Leigh Anderson, <u>Expression profiling in toxicology -- potentials and limitations</u>, Toxicology Letters 112-13:467-471 (2000).

4)   John C. Rockett and David J. Dix, <u>Application of DNA arrays to toxicology</u>, 107 Environ. Health Perspec. 107:681-685 (1999).

5)   Email from the primary investigator on the Nuwaysir paper, Dr. Cynthia Afshari, to an Incyte employee, dated July 3, 2000, as well as the original message to which she was responding.

6)   Page, M.J. et al., <u>Proteomics: a major new technology for the drug discovery process</u>, Drug Discov. Today 4:55-62 (1999).

7)   Declaration of Tod Bedilion, Ph.D., under 37 C.F.R. 1.132;

8)   Declaration of Vishwanath R. Iyer, Ph.D., under 37 C.F.R. 1.132 with Exhibits A - E; and

9)   ten (10) references published before the filing date of the instant application:

a) WO 95/21944, SmithKline Beecham, "Differentially expressed genes in healthy and diseased subjects" (Aug. 17, 1995)

b) WO 95/20681, Incyte Pharmaceuticals, "Comparative Gene Transcript Analysis" (Aug 3, 1995)

c) Schena et al., "Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray," Science 270:467-470 (Oct 20, 1995)

d) WO 95/35505, Stanford University, "Method and apparatus for fabricating microarrays of biological samples" (Dec 28, 1995)

e) U.S. Pat. No. 5,569,588, Ashby et al., "Methods for Drug Screening" (Oct 29, 1996)

f) Heller al., "Discovery and analysis of inflammatory disease-related genes using cDNA microarrays," *PNAS* 94:2150 - 2155 (Mar 1997)

g) WO 97/13877, Lynx Therapeutics, "Measurement of Gene Expression Profiles in Toxicity Determinations" (April 17, 1997)

h) Acacia Biosciences Press Release (August 11, 1997)

i) Glaser, "Strategies for Target Validation Streamline Evaluation of Leads," Genetic Engineering News (Sept. 15, 1997)

j) DeRisi *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," Science 278:680 - 686 (Oct 24, 1997)

APPENDIX - CLAIMS ON APPEAL

1.   An isolated cDNA comprising a nucleic acid encoding an amino acid sequence selected from:

   a) an amino acid sequence of SEQ ID NO:1;

   b) a fragment of SEQ ID NO:1 from I51-V72, G88-V109, C116-A145, I156-L175, M207-P229, or G242-T264 of SEQ ID NO:1;

   c) a variant of SEQ ID NO:1 having at least 90% amino acid sequence identity to SEQ ID NO:1; and

   d) the complement of the encoding nucleic acid sequence of a), b), or c).

2.   An isolated cDNA comprising a nucleic acid sequence selected from:

   a) SEQ ID NO:7; and

   b) a variant of SEQ ID NO:7 having at least 95% identity to SEQ ID NO:7.

3.   A composition comprising the cDNA of claim 1 and a labeling moiety.

4.   A vector comprising the cDNA of claim 1.

5.   A host cell comprising the vector of claim 4.

6.   A method for using a cDNA to produce a protein, the method comprising:

   a)   culturing the host cell of claim 5 under conditions for protein expression; and

   b)   recovering the protein from the host cell culture.

Whole genome analysis: Experimental access to all genome sequenced segments through larger-scale efficient oligonucleotide synthesis and PCR

DEVAL A. LASHKARI*†, JOHN H. MCCUSKER‡, AND RONALD W. DAVIS*§

*Departments of Genetics and Biochemistry, Beckman Center, Stanford University, Stanford, CA 94305; and ‡Department of Microbiology, 3020 Duke University Medical Center, Durham, NC 27710

Contributed by Ronald W. Davis, May 20, 1997

ABSTRACT   The recent ability to sequence whole genomes allows ready access to all genetic material. The approaches outlined here allow automated analysis of sequence for the synthesis of optimal primers in an automated multiplex oligonucleotide synthesizer (AMOS). The efficiency is such that all ORFs for an organism can be amplified by PCR. The resulting amplicons can be used directly in the construction of DNA arrays or can be cloned for a large variety of functional analyses. These tools allow a replacement of single-gene analysis with a highly efficient whole-genome analysis.

The genome sequencing projects have generated and will continue to generate enormous amounts of sequence data. The genomes of *Saccharomyces cerevisiae*, *Escherichia coli*, *Haemophilus influenzae* (1), *Mycoplasma genitalium* (2), and *Methanococcus jannaschii* (3) have been completely sequenced. Other model organisms have had substantial portions of their genomes sequenced as well, including the nematode *Caenorhabditis elegans* (4) and the small flowering plant *Arabidopsis thaliana* (5). This massive and increasing amount of sequence information allows the development of novel experimental approaches to identify gene function.

One standard use of genome sequence data is to attempt to identify the functions of predicted open reading frames (ORFs) within the genome by comparison to genes of known function. Such a comparative analysis of all ORFs to existing sequence data is fast, simple, and requires no experimentation and is therefore a reasonable first step. While finding sequence homologies/motifs is not a substitute for experimentation, noting the presence of sequence homology and/or sequence motifs can be a useful first step in finding interesting genes, in designing experiments and, in some cases, predicting function. However, this type of analysis is frequently uninformative. For example, over one-half of new ORFs in *S. cerevisiae* have no known function (6). If this is the case in a well studied organism such as yeast, the problem will be even worse in organisms that are less well studied or less manipulable. A large, experimentally determined gene function database would make homology/motif searches much more useful.

Experimental analysis must be performed to thoroughly understand the biological function of a gene product. Scaling up from classical "cottage industry" one-gene-oriented approaches to whole-genome analysis would be very expensive and laborious. It is clear that novel strategies are necessary to efficiently pursue the next phase of the genome projects—whole-genome experimental analysis to explore gene expression, gene product function, and other genome functions. Model organisms, such as *S. cerevisiae*, will be extremely important in the development of novel whole-genome analysis techniques and, subsequently, in improving our understanding of other more complex and less manipulable organisms.

The genome sequence can be systematically used as a tool to understand ORFs, gene product function, and other genome regions. Toward this end, a directed strategy has been developed for exploiting sequence information as a means of providing information about biological function (Fig. 1). Efforts have been directed toward the amplification of each predicted ORF or any other region of the genome ranging from a few base pairs to several kilobase pairs. There are many uses for these amplicons—they can be cloned into standard vectors or specialized expression vectors, or can be cloned into other specialized vectors such as those used for two-hybrid analysis. The amplicons can also be used directly by, for example, arraying onto glass for expression analysis, for DNA binding assays, or for any direct DNA assay (7). As a pilot study, synthetic primers were made on the 96-well automated multiplex oligonucleotide synthesizer (AMOS) instrument (8) (Fig. 2). These oligonucleotides were used to amplify each ORF on yeast chromosome V. The current version of this instrument can synthesize three plates of 96 oligonucleotides each (25 bases) in an 8-hr day. The amplification of the entire set of PCR products was then analyzed by gel electrophoresis (Fig. 3). Successful amplification of the proper length product on the first attempt was 95%. This project demonstrates that one can go directly from sequence information to biological analysis in a truly automated, totally directed manner.

These amplicons can be incorporated directly in arrays or the amplicons can be cloned. If the amplicons are to be cloned, novel sequences can be incorporated at the 5' end of the oligonucleotide to facilitate cloning. One potential problem with cloning PCR products is that the cloned amplicons may contain sequence alterations that diminish their utility. One option would be to resequence each individual amplicon. However, this is expensive, inefficient, and time consuming. A faster, more cost-effective, and more accurate approach is to apply comparative sequencing by denaturing HPLC (9). This method is capable of detecting a single base change in a 2-kb heteroduplex. Longer amplicons can be analyzed by use of appropriate restriction fragments. If any change is detected in a clone, an alternate clone of the same region can be analyzed. Modifying the system to allow high throughput analysis by denaturing HPLC is also relatively simple and straightforward.

If amplicons are used directly on arrays without cloning, it is important to note that, even if single PCR product bands are observed on gels, the PCR products will be contaminated with various amounts of other sequences. This contamination has the potential to affect the results in, for example, expression

†Present address: Synteni, Inc., 6519 Dumbarton Circle, Fremont, CA 94555.
§To whom reprint requests should be addressed at: Department of Biochemistry, Beckman Center, B400, Stanford University, Stanford, CA 94305-5307. e-mail: gilbert@cmgm.stanford.edu.
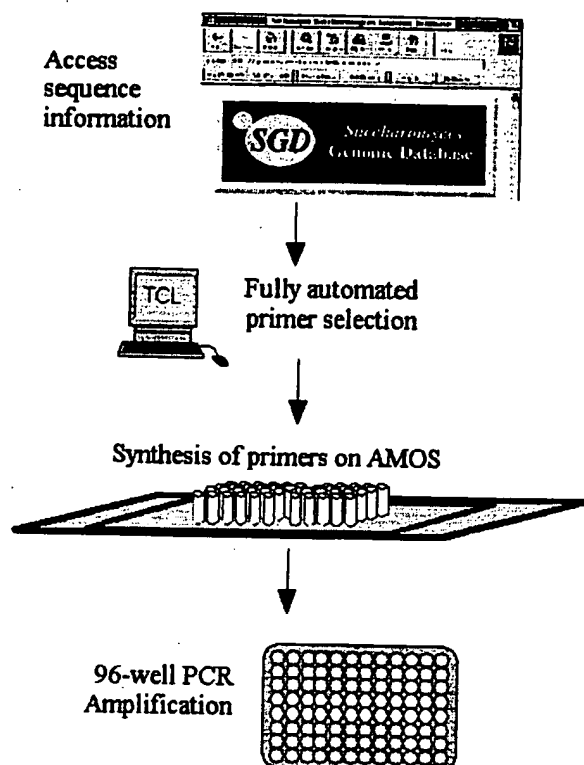
Fig. 1.   Overview of systematic method for isolating individual genes. Sequence information is obtained automatically from sequence databases. The data are input into primer selection software specifically designed to target ORFs as designated by database annotations. The output file containing the primer information is directly read by a high-throughput oligonucleotide synthesizer, which makes the oligonucleotides in 96-well plates (AMOS, automated multiplex oligonucleotide synthesizer). The forward and reverse primers are synthesized in the same location on separate plates to facilitate the downstream handling of primers. The amplicons are generated by PCR in 96-well plates as well.

Fig. 2.   Overall approach for using database of a genome to direct biological analysis. The synthesis of the 6,000 ORFs (orfs) for each gene of *S. cerevisiae* can be used in many applications utilizing both cloning and microarraying technology.

analysis. On the other hand, direct use of the amplicons is much less labor intensive and greatly decreases the occurrence of mistakes in clone identification, a ubiquitous problem associated with large clone set archiving and retrieving.

Any large-scale effort to capture each ORF within a genome must rely on automation if cost is to be minimized while efficiency is maximized. Toward that end, primers targeting ORFs were designed automatically using simple new scripts and existing primer selection software. These script-selected primer sequences were directly read by the high-throughput synthesizer and the forward and reverse primers were synthesized in separate plates in corresponding wells to facilitate automated pipetting and PCR amplifications. Each of the resulting PCR products, generated with minimum labor, contains a known, unique ORF.

Large-scale genome analysis projects are dependent on newly emerging technologies to make the studies practical and economically feasible. For example, the cost of the primers, a significant issue in the past, has been reduced dramatically to make feasible this and other projects that require tens of thousands of oligonucleotides. Other methods of high-throughput analysis are also vital to the success of functional analysis projects, such as microarraying and oligonucleotide chip methods (10–14).

Changes in attitude are also required. One of the major costs of commercial oligonucleotides is extensive quality control such that virtually 100% of the supplied oligonucleotides are successfully synthesized and work for their intended purpose.

Considerable cost reduction can be obtained by simply decreasing the expected successful synthesis rate to 95–97%. One can then achieve faster and cheaper whole genome coverage by simply adding a single quality control at the end of the experiment and batching the failures for resynthesis.

The directed nature of the amplicon approach is of clear advantage. The sequence of each ORF is analyzed automatically, and unique specific primers are made to target each ORF. Thus, there is relatively little time or labor involved—for example, no random cloning and subsequent screening is required because each product is known. In the test system, primers for 240 ORFs from chromosome V were systematically synthesized, beginning from the left arm and continuing through to the right arm. At no point was there any manual analysis of sequence information to generate the collection. In many ways, now that the sequence is known, there is no need for the researcher to examine it.

These amplicons can be arrayed and expression analysis can be done on all arrayed ORFs with a single hybridization (10). Those ORFs that display significant differential expression patterns under a given selection are easily identified without the laborious task of searching for and then sequencing a clone. Once scaled up, the procedure provides even greater returns on effort, because a single hybridization will ultimately provide a "snapshot" of the expression of all genes in the yeast genome. Thus, the limiting factor in whole genome analysis will not be the analysis process itself, but will instead be the ability of researchers to design and carry out experimental selections.

Current expression and genetic analysis technologies are geared toward the analysis of single genes and are ill suited to analyze numerous genes under many conditions. Additional difficulties with current technologies include: the effort and expense required to analyze expression and make mutants, the potential duplication of effort if done by different laboratories, and the possibility of conflicting results obtained from different laboratories. In contrast, whole genome analysis not only is more efficient, it also provides data of much higher quality; all genes are assayed and compared in parallel under exactly

Applied Biological Sciences: Lashkari *et al.*
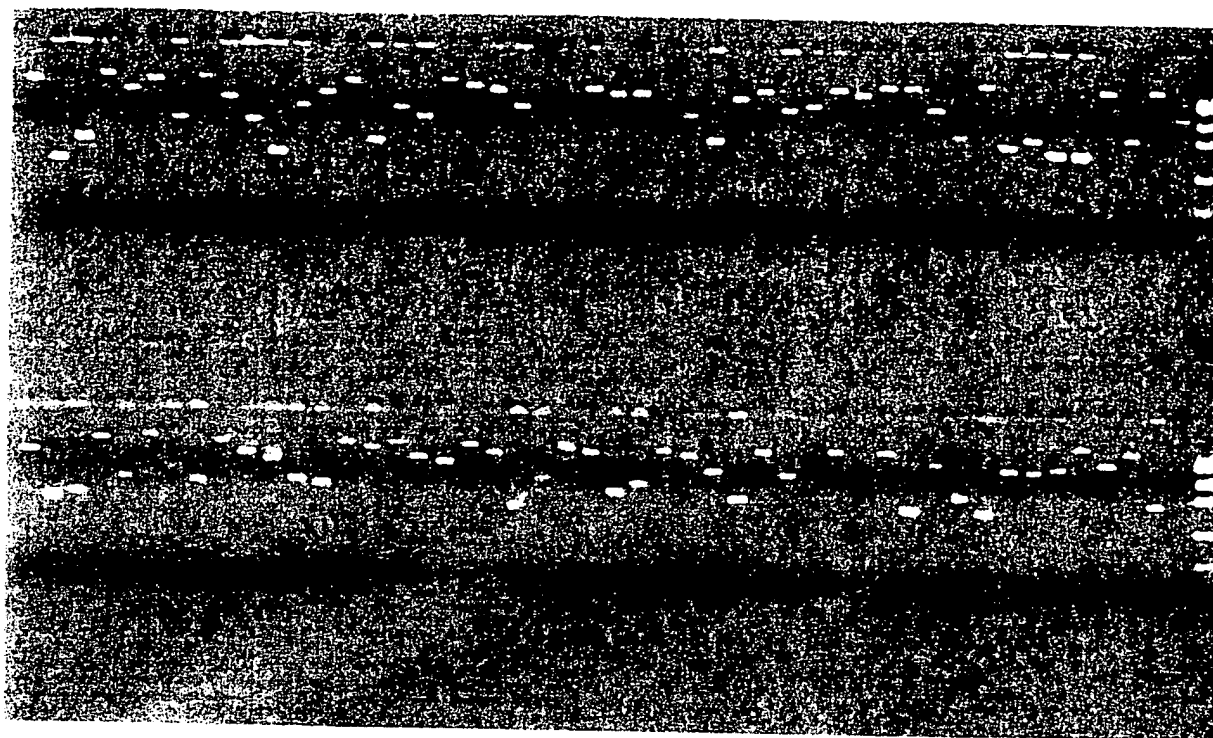
Proc. Natl. Acad. Sci. USA 94 (1997) 8947

Fig. 3. Gel image of amplifications. Using the method described in Fig. 1, amplicons were generated for ORFs of *S. cerevisiae* chromosome V. One plate of 96 amplification reactions is shown.

the same conditions. In addition, amplicons have many applications beyond gene expression. For example, one recent approach is to incorporate a unique DNA sequence tag, synthesized as part of each gene specific primer, during amplification. The tags or molecular bar codes, when reintroduced into the organism as a gene deletion or as a gene clone, can be used much more efficiently than individual mutations or clones because pools of tagged mutants or transformants can be analyzed in parallel. This parallel analysis is possible because the tags are readily and quantitatively amplified even in complex mixtures of tags (13).

These ORF genome arrays and oligonucleotide tagged libraries can be used for many applications. Any conventional selection applied to a library that gives discrete or multiple products can use these technologies for a simple direct readout. These include screens and selections for mutant complementation, overexpression suppression (15, 16), second-site suppressors, synthetic lethality, drug target overexpression (17), two-hybrid screens (18), genome mismatch scanning (19), or recombination mapping.

The genome projects have provided researchers with a vast amount of information. These data must be used efficiently and systematically to gain a truly comprehensive understanding of gene function and, more broadly, of the entire genome which can then be applied to other organisms. Such global approaches are essential if we are to gain an understanding of the living cell. This understanding should come from the viewpoint of the integration of complex regulatory networks, the individual roles and interactions of thousands of functional gene products, and the effect of environmental changes on both gene regulatory networks and the roles of all gene products. The time has come to switch from the analysis of a single gene to the analysis of the whole genome.

1. Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., *et al.* (1995) *Science* **269**, 496–512.
2. Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., *et al.* (1995) *Science* **270**, 397–403.
3. Bult, C. J., White, O., Olsen, G. J., Zhou, L., Fleischmann, R. D., *et al.* (1996) *Science* **273**, 1058–1073.
4. Sulston, J., Du, Z., Thomas, K., Wilson, R., Hillier, L., Staden, R., Halloran, N., Green, P., Thierry-Mieg, J., Qiu, L., Dear, S., Coulson, A., Craxton, M., Durbin, R., Berks, M., Metzstein, M., Hawkins, T., Ainscough, R. & Waterston, R. (1992) *Nature (London)* **356**, 37–41.
5. Newman, T., de Bruijn, F. J., Green, P., Keegstra, K., Kende, H., *et al.* (1994) *Plant Physiol.* **106**, 1241–1255.
6. Oliver, S. (1996) *Nature (London)* **379**, 597–600.
7. Lashkari, D. A. (1996) Ph.D. dissertation (Stanford Univ., Stanford, CA).
8. Lashkari, D. A., Hunicke-Smith, S. P., Norgren, R. M., Davis, R. W. & Brennan, T. (1995) *Proc. Natl. Acad. Sci. USA* **92**, 7912–7915.
9. Oefner, P. J. & Underhill, P. A. (1995) *Am. J. Hum. Genet.* **57**, A266.
10. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
11. Fodor, S. P., Read, J. L., Pirrung, M. C., Stryer, L., Lu, A. T. & Solas, D. (1991) *Science* **251**, 767–773.
12. Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X. C., Stern, D., Winkler, J., Lockhart, D. J., Morris, M. S. & Fodor, S. P. (1996) *Science* **274**, 610–614.
13. Shoemaker, D. D., Lashkari, D. A., Morris, D., Mittmann, M. & Davis, R. W. (1996) *Nat. Genet.* **14**, 450–456.
14. Smith, V., Chou, K., Lashkari, D., Botstein, D. & Brown, P. O. (1996) *Science* **274**, 2069–2074.
15. Magdolen, V., Drubin, D. G., Mages, G. & Bandlow, W. (1993) *FEBS Lett.* **316**, 41–47.
16. Ramer, S. W., Elledge, S. J. & Davis, R. W. (1992) *Proc. Natl. Acad. Sci. USA* **89**, 11589–11593.
17. Rine, J., Hansen, W., Hardeman, E. & Davis, R. W. (1983) *Proc. Natl. Acad. Sci. USA* **80**, 6750–6754.
18. Fields, S. & Song, O. (1989) *Nature (London)* **340**, 245–246.
19. Nelson, S. F., McCusker, J. H., Sander, M. A., Kee, Y., Modrich, P. & Brown, P. O. (1994) *Nat. Genet.* **4**, 11–18.

IN PERSPECTIVE

Microarrays and Toxicology: The Advent of Toxicogenomics

Emile F. Nuwaysir,[1] Michael Bittner,[2] Jeffrey Trent,[2] J. Carl Barrett,[1] and Cynthia A. Afshari[1]

[1]Laboratory of Molecular Carcinogenesis, National Institute of Environmental Health Sciences, Research Triangle Park, North Carolina
[2]Laboratory of Cancer Genetics, National Human Genome Research Institute, Bethesda, Maryland

The availability of genome-scale DNA sequence information and reagents has radically altered life-science research. This revolution has led to the development of a new scientific subdiscipline derived from a combination of the fields of toxicology and genomics. This subdiscipline, termed toxicogenomics, is concerned with the identification of potential human and environmental toxicants, and their putative mechanisms of action, through the use of genomics resources. One such resource is DNA microarrays or "chips," which allow the monitoring of the expression levels of thousands of genes simultaneously. Here we propose a general method by which gene expression, as measured by cDNA microarrays, can be used as a highly sensitive and informative marker for toxicity. Our purpose is to acquaint the reader with the development and current state of microarray technology and to present our view of the usefulness of microarrays to the field of toxicology. Mol. Carcinog. 24:153–159, 1999.   © 1999 Wiley-Liss, Inc.

Key words: toxicology; gene expression; animal bioassay

INTRODUCTION

Technological advancements combined with intensive DNA sequencing efforts have generated an enormous database of sequence information over the past decade. To date, more than 3 million sequences, totaling over 2.2 billion bases [1], are contained within the GenBank database, which includes the complete sequences of 19 different organisms [2]. The first complete sequence of a free-living organism, Haemophilus influenzae, was reported in 1995 [3] and was followed shortly thereafter by the first complete sequence of a eukaryote, Saccharomyces cervisiae [4]. The development of dramatically improved sequencing methodologies promises that complete elucidation of the Homo sapiens DNA sequence is not far behind [5].

To exploit more fully the wealth of new sequence information, it was necessary to develop novel methods for the high-throughput or parallel monitoring of gene expression. Established methods such as northern blotting, RNAse protection assays, S1 nuclease analysis, plaque hybridization, and slot blots do not provide sufficient throughput to effectively utilize the new genomics resources. Newer methods such as differential display [6], high-density filter hybridization [7,8], serial analysis of gene expression [9], and cDNA- and oligonucleotide-based microarray "chip" hybridization [10–12] are possible solutions to this bottleneck. It is our belief that the microarray approach, which allows the monitoring of expression levels of thousands of genes simultaneously, is a tool of unprecedented power for use in toxicology studies.

Almost without exception, gene expression is altered during toxicity, as either a direct or indirect result of toxicant exposure. The challenge facing toxicologists is to define, under a given set of experimental conditions, the characteristic and specific pattern of gene expression elicited by a given toxicant. Microarray technology offers an ideal platform for this type of analysis and could be the foundation for a fundamentally new approach to toxicology testing.

MICROARRAY DEVELOPMENT AND APPLICATIONS

cDNA Microarrays

In the past several years, numerous systems were developed for the construction of large-scale DNA arrays. All of these platforms are based on cDNAs or oligonucleotides immobilized to a solid support. In the cDNA approach, cDNA (or genomic) clones of interest are arrayed in a multi-well format and amplified by polymerase chain reaction. The products of this amplification, which are usually 500- to 2000-bp clones from the 3' regions of the genes of interest, are then spotted onto solid support by using high-speed robotics. By using this method, microarrays of up to 10 000 clones can be generated by spotting onto a glass substrate

[13,14]. Sample detection for microarrays on glass involves the use of probes labeled with fluorescent or radioactive nucleotides.

Fluorescent cDNA probes are generated from control and test RNA samples in single-round reverse-transcription reactions in the presence of fluorescently tagged dUTP (e.g., Cy3-dUTP and Cy5-dUTP), which produces control and test products labeled with different fluors. The cDNAs generated from these two populations, collectively termed the "probe," are then mixed and hybridized to the array under a glass coverslip [10,11,15]. The fluorescent signal is detected by using a custom-designed scanning confocal microscope equipped with a motorized stage and lasers for fluor excitation [10,11,15]. The data are analyzed with custom digital image analysis software that determines for each DNA feature the ratio of fluor 1 to fluor 2, corrected for local background [16,17]. The strength of this approach lies in the ability to label RNAs from control and treated samples with different fluorescent nucleotides, allowing for the simultaneous hybridization and detection of both populations on one microarray. This method eliminates the need to control for hybridization between arrays. The research groups of Drs. Patrick Brown and Ron Davis at Stanford University spearheaded the effort to develop this approach, which has been successfully applied to studies of *Arabidopsis thaliana* RNA [10], yeast genomic DNA [15], tumorigenic versus non-tumorigenic human tumor cell lines [11], human T-cells [18], yeast RNA [19], and human inflammatory disease–related genes [20]. The most dramatic result of this effort was the first published account of gene expression of an entire genome, that of the yeast *Saccharomyces cervisiae* [21].

In an alternative approach, large numbers of cDNA clones can be spotted onto a membrane support, albeit at a lower density [7,22]. This method is useful for expression profiling and large-scale screening and mapping of genomic or cDNA clones [7,22–24]. In expression profiling on filter membranes, two different membranes are used simultaneously for control and test RNA hybridizations, or a single membrane is stripped and reprobed. The signal is detected by using radioactive nucleotides and visualized by phosphorimager analysis or autoradiography. Numerous companies now sell such cDNA membranes and software to analyze the image data [25–27].

Oligonucleotide Microarrays

Oligonucleotide microarrays are constructed either by spotting prefabricated oligos on a glass support [13] or by the more elegant method of direct in situ oligo synthesis on the glass surface by photolithography [28–30]. The strength of this approach lies in its ability to discriminate DNA molecules based on single base-pair difference. This allows the application of this method to the fields of medical diagnos-

tics, pharmacogenetics, and sequencing by hybridization as well as gene-expression analysis.

Fabrication of oligonucleotide chips by photolithography is theoretically simple but technically complex [29,30]. The light from a high-intensity mercury lamp is directed through a photolithographic mask onto the silica surface, resulting in deprotection of the terminal nucleotides in the illuminated regions. The entire chip is then reacted with the desired free nucleotide, resulting in selected chain elongation. This process requires only 4n cycles (where n = oligonucleotide length in bases) to synthesize a vast number of unique oligos, the total number of which is limited only by the complexity of the photolithographic mask and the chip size [29,31,32].

Sample preparation involves the generation of double-stranded cDNA from cellular poly(A)+ RNA followed by antisense RNA synthesis in an in vitro transcription reaction with biotinylated or fluor-tagged nucleotides. The RNA probe is then fragmented to facilitate hybridization. If the indirect visualization method is used, the chips are incubated with fluor-linked streptavidin (e.g., phycoerythrin) after hybridization [12,33]. The signal is detected with a custom confocal scanner [34]. This method has been applied successfully to the mapping of genomic library clones [35], to de novo sequencing by hybridization [28,36], and to evolutionary sequence comparison of the *BRCA1* gene [37]. In addition, mutations in the cystic fibrosis [38] and BRCA1 [39] gene products and polymorphisms in the human immunodeficiency virus-1 clade B protease gene [40] have been detected by this method. Oligonucleotide chips are also useful for expression monitoring [33] as has been demonstrated by the simultaneous evaluation of gene-expression patterns in nearly all open reading frames of the yeast strain *S. cerevisiae* [12]. More recently, oligonucleotide chips have been used to help identify single nucleotide polymorphisms in the human [41] and yeast [42] genomes.

THE USE OF MICROARRAYS IN TOXICOLOGY

Screening for Mechanism of Action

The field of toxicology uses numerous in vivo model systems, including the rat, mouse, and rabbit, to assess potential toxicity and these bioassays are the mainstay of toxicology testing. However, in the past several decades, a plethora of in vitro techniques have been developed to measure toxicity, many of which measure toxicant-induced DNA damage. Examples of these assays include the Ames test, the Syrian hamster embryo cell transformation assay, micronucleus assays, measurements of sister chromatid exchange and unscheduled DNA synthesis, and many others. Fundamental to all of these methods is the fact that toxicity is often preceded by, and results in, alterations in gene expression. In many cases, these changes in gene expression are a

far more sensitive, characteristic, and measurable endpoint than the toxicity itself. We therefore propose that a method based on measurements of the genome-wide gene expression pattern of an organism after toxicant exposure is fundamentally informative and complements the established methods described above.

We are developing a method by which toxicants can be identified and their putative mechanisms of action determined by using toxicant-induced gene expression profiles. In this method, in one or more defined model systems, dose and time-course parameters are established for a series of toxicants within a given prototypic class (e.g., polycyclic aromatic hydrocarbons (PAHs)). Cells are then treated with these agents at a fixed toxicity level (as measured by cell survival), RNA is harvested, and toxicant-induced gene expression changes are assessed by hybridization to a cDNA microarray chip (Figure 1). We have developed a custom DNA chip, called ToxChip v1.0, specifically for this purpose and will discuss it in more detail below. The changes in gene expression induced by the test agents in the model systems are analyzed, and the common set of changes unique to that class of toxicants, termed a toxicant signature, is determined.

This signature is derived by ranking across all experiments the gene-expression data based on rela-

tive fold induction or suppression of genes in treated samples versus untreated controls and selecting the most consistently different signals across the sample set. A different signature may be established for each prototypic toxicant class. Once the signatures are determined, gene-expression profiles induced by unknown agents in these same model systems can then be compared with the established signatures. A match assigns a putative mechanism of action to the test compound. Figure 2 illustrates this signature method for different types of oxidant stressors, PAHs, and peroxisome proliferators. In this example, the unknown compound in question had a gene-expression profile similar to that of the oxidant stressors in the database. We anticipate that this general method will also reveal cross talk between different pathways induced by a single agent (e.g., reveal that a compound has both PAH-like and oxidant-like properties). In the future, it may be necessary to distinguish very subtle differences between compounds within a very large sample set (e.g., thousands of highly similar structural isomers in a combinatorial chemistry library or peptide library). To generate these highly refined signatures, standard statistical clustering techniques or principal-component analysis can be used.

For the studies outlined in Figure 2, we developed the custom cDNA microarray chip ToxChip v1.0.
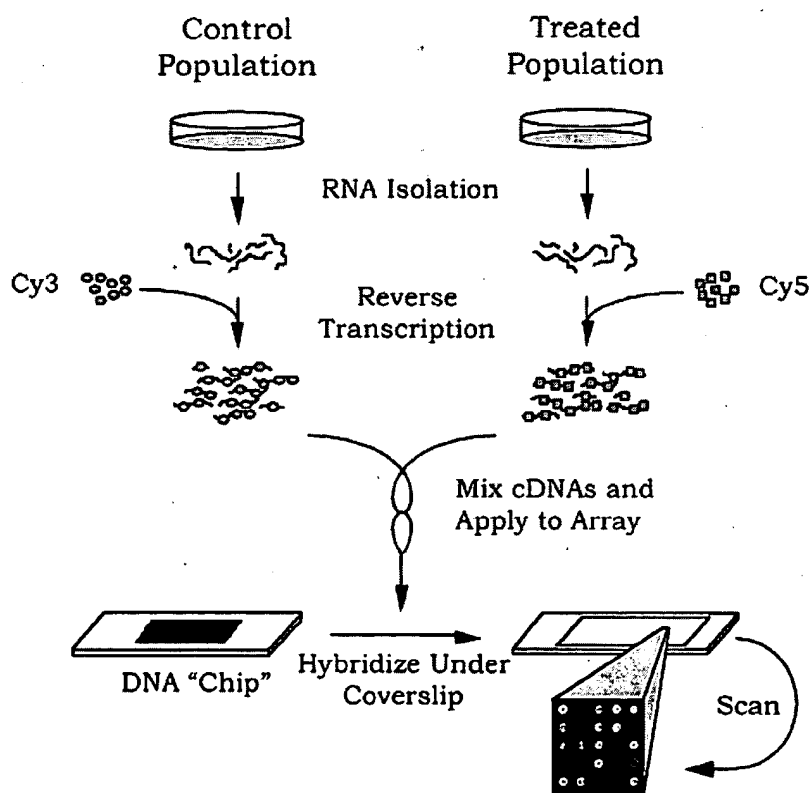

Figure 1. Simplified overview of the method for sample preparation and hybridization to cDNA microarrays. For illustrative purposes, samples derived from cell culture are depicted, although other sample types are amenable to this analysis.
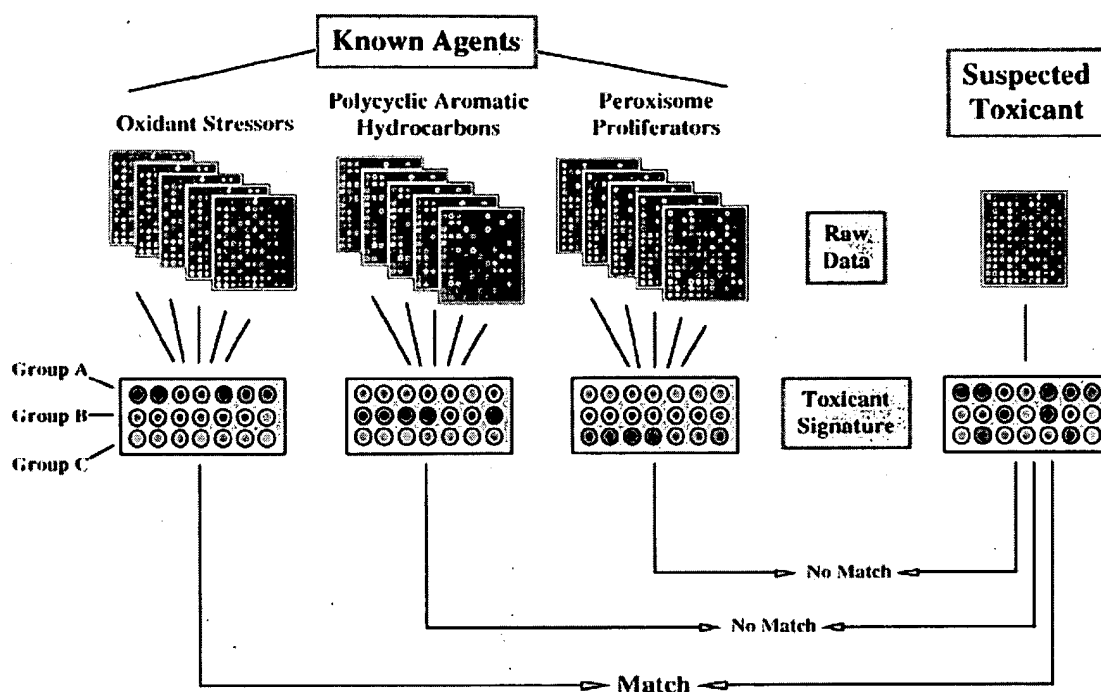
Figure 2. Schematic representation of the method for identification of a toxicant's mechanism of action. In this method, gene-expression data derived from exposure of model systems to known toxicants are analyzed, and a set of changes characteristic to that type of toxicant (termed the toxicant signature) is identified. As depicted, oxidant stressors produce consistent changes in group A genes (indicated by red and green circles), but not group B or C genes (indicated by gray circles). The set of gene-expression changes elicited by the suspected toxicant is then compared with these characteristic patterns, and a putative mechanism of action is assigned to the unknown agent.

The 2090 human genes that comprise this subarray were selected for their well-documented involvement in basic cellular processes as well as their responses to different types of toxic insult. Included on this list are DNA replication and repair genes, apoptosis genes, and genes responsive to PAHs and dioxin-like compounds, peroxisome proliferators, estrogenic compounds, and oxidant stress. Some of the other categories of genes include transcription factors, oncogenes, tumor suppressor genes, cyclins, kinases, phosphatases, cell adhesion and motility genes, and homeobox genes. Also included in this group are 84 housekeeping genes, whose hybridization intensity is averaged and used for signal normalization of the other genes on the chip. To date, very few toxicants have been shown to have appreciable effects on the expression of these housekeeping genes. However, this housekeeping list will be revised if new data warrant the addition or deletion of a particular gene. Table 1 contains a general description of some of the different classes of genes that comprise ToxChip v1.0.

When a toxicant signature is determined, the genes within this signature are flagged within the database. When uncharacterized toxicants are then screened, the data can be quickly reformatted so that blocks of genes representing the different signatures are displayed [11]. This facilitates rapid, visual interpretation of data. We are also developing Tox-Chip v2.0 and chips for other model systems, including rat, mouse, Xenopus, and yeast, for use in toxicology studies.

Animal Models in Toxicology Testing

The toxicology community relies heavily on the use of animals as model systems for toxicology testing. Unfortunately, these assays are inherently expensive, require large numbers of animals and take a long time to complete and analyze. Therefore, the National Institute of Environmental Health Sciences (NIEHS), the National Toxicology Program, and the toxicology community at large are committed to reducing the number of animals used, by developing more efficient and alternative testing methodologies. Although substantial progress has been made in the development of alternative methods, bioassays are still used for testing endpoints such as neurotoxicity, immunotoxicity, reproductive and developmental toxicology, and genetic toxicology. The rodent cancer bioassay is a particularly expensive and time-consuming assay, as it requires almost 4 yr, 1200 animals, and millions of dollars to execute and analyze [43]. In vitro experiments of the type outlined in Figure 2 might provide evidence that an unknown

Table 1. ToxChip v1.0: A Human cDNA Microarray Chip Designed to Detect Responses to Toxic Insult

Gene category	No. of genes on chip
Apoptosis	72
DNA replication and repair	99
Oxidative stress/redox homeostasis	90
Peroxisome proliferator responsive	22
Dioxin/PAH responsive	12
Estrogen responsive	63
Housekeeping	84
Oncogenes and tumor suppressor genes	76
Cell-cycle control	51
Transcription factors	131
Kinases	276
Phosphatases	88
Heat-shock proteins	23
Receptors	349
Cytochrome P450s	30

*This list is intended as a general guide. The gene categories are not unique, and some genes are listed in multiple categories.

agent is (or is not) responsible for eliciting a given biological response. This information would help to select a bioassay more specifically suited to the agent in question or perhaps suggest that a bioassay is not necessary, which would dramatically reduce cost, animal use, and time.

The addition of microarray techniques to standard bioassays may dramatically enhance the sensitivity and interpretability of the bioassay and possibly reduce its cost. Gene-expression signatures could be determined for various types of tissue-specific toxicants, and new compounds could be screened for these characteristic signatures, providing a rapid and sensitive in vivo test. Also, because gene expression is often exquisitely sensitive to low doses of a toxicant, the combination of gene-expression screening and the bioassay might allow the use of lower toxicant doses, which are more relevant to human exposure levels, and the use of fewer animals. In addition, gene-expression changes are normally measured in hours or days, not in the months to years required for tumor development. Furthermore, microarrays might be particularly useful for investigating the relationship between acute and chronic toxicity and identifying secondary effects of a given toxicant by studying the relationship between the duration of exposure to a toxicant and the gene-expression profile produced. Thus, a bioassay that incorporates gene-expression signatures with traditional endpoints might be substantially shorter, use more realistic dose regimens, and cost substantially less than the current assays do.

These considerations are also relevant for branches of toxicology not related to human health and not using rodents as model systems, such as aquatic toxicology and plant pathology. Bioassays based on the flathead minnow, *Daphnia,* and *Arabadopsis* could

also be improved by the addition of microarray analysis. The combination of microarrays with traditional bioassays might also be useful for investigating some of the more intractable problems in toxicology research, such as the effects of complex mixtures and the difficulties in cross-species extrapolation.

Exposure Assessment, Environmental Monitoring, and Drug Safety

The currently used methods for assessment of exposure to chemical toxicants are based on measurement of tissue toxin levels or on surrogate markers of toxicity, termed biomarkers (e.g., peripheral blood levels of hepatic enzymes or DNA adducts). Because gene expression is a sensitive endpoint, gene expression as measured with microarray technology may be useful as a new biomarker to more precisely identify hazards and to assess exposure. Similarly, microarrays could be used in an environmental-monitoring capacity to measure the effect of potential contaminants on the gene-expression profiles of resident organisms. In an analogous fashion, microarrays could be used to measure gene-expression endpoints in subjects in clinical trials. The combination of these gene-expression data and more established toxic endpoints in these trials could be used to define highly precise surrogates of safety.

Gene-expression profiles in samples from exposed individuals could be compared to the profiles of the same individuals before exposure. From this information, the nature of the toxic exposure can be determined or a relative clinical safety factor estimated. In the future it may also be possible to estimate not only the nature but the dose of the toxicant for a given exposure, based on relative gene-expression levels. This general approach may be particularly appropriate for occupational-health applications, in which unexposed and exposed samples from the same individuals may be obtainable. For example, a pilot study of gene expression in peripheral-blood lymphocytes of Polish coke-oven workers exposed to PAHs (and many other compounds) is under consideration at the NIEHS. An important consideration for these types of studies is that gene expression can be affected by numerous factors, including diet, health, and personal habits. To reduce the effects of these confounding factors, it may be necessary to compare pools of control samples with pools of treated samples. In the future it may be possible to compare exposed sample sets to a national database of human-expression data, thus eliminating the need to provide an unexposed sample from the same individual. Efforts to develop such a national gene-expression database are currently under way [44,45]. However, this national database approach will require a better understanding of genome-wide gene expression across the highly diverse human population and of the effects of environmental factors on this expression.

Alleles, Oligo Arrays, and Toxicogenetics

Gene sequences vary between individuals, and this variability can be a causative factor in human diseases of environmental origin [46,47]. A new area of toxicology, termed toxicogenetics, was recently developed to study the relationship between genetic variability and toxicant susceptibility. This field is not the subject of this discussion, but it is worthwhile to note that the ability of oligonucleotide arrays to discriminate DNA molecules based on single base-pair differences makes these arrays uniquely useful for this type of analysis. Recent reports demonstrated the feasibility of this approach [41,42]. The NIEHS has initiated the Environmental Genome Project to identify common sequence polymorphisms in 200 genes thought to be involved in environmental diseases [48]. In a pilot study on the feasibility of this application to the Environmental Genome Project, oligonucleotide arrays will be used to resequence 20 candidate genes. This toxicogenetic approach promises to dramatically improve our understanding of interindividual variability in disease susceptibility.

FUTURE PRIORITIES

There are many issues that must be addressed before the full potential of microarrays in toxicology research can be realized. Among these are model system selection, dose selection, and the temporal nature of gene expression. In other words, in which species, at what dose, and at what time do we look for toxicant-induced gene expression? If human samples are analyzed, how variable is global gene expression between individuals, before and after toxicant exposure? What are the effects of age, diet, and other factors on this expression? Experience, in the form of large data sets of toxicant exposures, will answer these questions.

One of the most pressing issues for array scientists is the construction of a national public database (linked to the existing public databases) to serve as a repository for gene-expression data. This relational database must be made available for public use, and researchers must be encouraged to submit their expression data so that others may view and query the information. Researchers at the National Institutes of Health have made laudable progress in developing the first generation of such a database [44,45]. In addition, improved statistical methods for gene clustering and pattern recognition are needed to analyze the data in such a public database.

The proliferation of different platforms and methods for microarray hybridizations will improve sample handling and data collection and analysis and reduce costs. However, the variety of microarray methods available will create problems of data compatibility between platforms. In addition, the near-infinite variety of experimental conditions under which data will be collected by different laboratories will make large-scale data analysis extremely difficult. To help circumvent these future problems, a set of standards to be included on all platforms should be established. These standards would facilitate data entry into the national database and serve as reference points for cross-platform and inter-laboratory data analysis.

Many issues remain to be resolved, but it is clear that new molecular techniques such as microarray hybridization will have a dramatic impact on toxicology research. In the future, the information gathered from microarray-based hybridization experiments will form the basis for an improved method to assess the impact of chemicals on human and environmental health.

ACKNOWLEDGMENTS

REFERENCES

1. http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html
2. http://www.ncbi.nlm.nih.gov/Entrez/Genome/org.html
3. Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 1995;269:496–512.
4. Goffeau A, Barrell BG, Bussey H, et al. Life with 6000 genes. Science 1996;274:546, 563–567.
5. http://www.perkin-elmer.com/press/prc5448.html
6. Liang P, Pardee AB. Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science 1992;257:967–971.
7. Pietu G, Alibert O, Guichard V, et al. Novel gene transcripts preferentially expressed in human muscles revealed by quantitative hybridization of a high density cDNA array. Genome Res 1996;6:492–503.
8. Zhao ND, Hashida H, Takahashi N, Misumi Y, Sakaki Y. High-density cDNA filter analysis—A novel approach for large-scale, quantitative analysis of gene expression. Gene 1995;156:207–213.
9. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW. Serial analysis of gene expression. Science 1995;270:484–487.
10. Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene-expression patterns with a complementary DNA microarray. Science 1995;270:467–470.
11. DeRisi J, Penland L, Brown PO, et al. use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 1996;14:457–460.
12. Wodicka L, Dong HL, Mittmann M, Ho MH, Lockhart DJ. Genome-wide expression monitoring in Saccharomyces cerevisiae. Nat Biotechnol 1997;15:1359–1367.
13. Marshall A, Hodgson J. DNA chips: An array of possibilities. Nat Biotechnol 1998;16:27–31.
14. http://www.synteni.com
15. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res 1996;6:639–645.
16. Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. Biomedical Optics 1997;2:364–374.
17. Khan J, Simon R, Bittner M, et al. Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. Cancer Res 1998;58:5009–5013.
18. Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: Microarray-based expression monitoring of 1000 genes. Proc Natl Acad Sci USA 1996; 93:10614–10619.

19. Lashkari DA, DeRisi JL, McCusker JH, et al. Yeast microarrays for genome wide parallel genetic and gene expression analysis. Proc Natl Acad Sci USA 1997;94:13057–13062.
20. Heller RA, Schena M, Chai A, et al. Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. Proc Natl Acad Sci USA 1997;94:2150–2155.
21. DeRisi JL, Iyer VR, Brown PO. Exploring the metabolic and genetic control of gene expression on a genomic scale. Science 1997;278:680–686.
22. Drmanac S, Stavropoulos NA, Labat I, et al. Gene-representing cDNA clusters defined by hybridization of 57,419 clones from infant brain libraries with short oligonucleotide probes. Genomics 1996;37:29–40.
23. Milosavljevic A, Savkovic S, Crkvenjakov R, et al. DNA sequence recognition by hybridization to short oligomers: Experimental verification of the method on the E. coli genome. Genomics 1996;37:77–86.
24. Drmanac S, Drmanac R. Processing of cDNA and genomic kilobase-size clones for massive screening, mapping and sequencing by hybridization. Biotechniques 1994;17:328–329, 332–336.
25. http://www.resgen.com/
26. http://www.genomesystems.com/
27. http://www.clontech.com/
28. Pease AC, Solas DA, Fodor SPA. Parallel synthesis of spatially addressable oligonucleotide probe matrices. Abstract. Abstracts of Papers of the American Chemical Society 1992;203:34.
29. Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. Proc Natl Acad Sci USA 1994;91:5022–5026.
30. Fodor SPA, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. Science 1991;251:767–773.
31. McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. Proc Natl Acad Sci USA 1996;93:13555–13560.
32. Lipshutz RJ, Morris D, Chee M, et al. Using oligonucleotide probe arrays to access genetic diversity. Biotechniques 1995;19:442–447.
33. Lockhart DJ, Dong HL, Byrne MC, et al. Expression monitoring by hybridization to high-density oligonucleotide arrays. Nat Biotechnol 1996;14:1675–1680.
34. http://www.mdyn.com/
35. Sapolsky RJ, Lipshutz RJ. Mapping genomic library clones using oligonucleotide arrays. Genomics 1996;33:445–456.
36. Chee M, Yang R, Hubbell E, et al. Accessing genetic information with high-density DNA arrays. Science 1996;274:610–614.
37. Hacia JG, Makalowski W, Edgemon K, et al. Evolutionary sequence comparisons using high-density oligonucleotide arrays. Nat Genet 1998;18:155–158.
38. Cronin MT, Fucini RV, Kim SM, Masino RS, Wespi RM, Miyada CG. Cystic fibrosis mutation detection by hybridization to light-generated DNA probe arrays. Hum Mutat 1996;7:244–255.
39. Hacia JG, Brody LC, Chee MS, Fodor SPA, Collins FS. Detection of heterozygous mutations in BRCA1 using high density oligonucleotide arrays and two-colour fluorescence analysis. Nat Genet 1996;14:441–447.
40. Kozal MJ, Shah N, Shen NP, et al. Extensive polymorphisms observed in HIV-1 clade B protease gene using high-density oligonucleotide arrays. Nat Med 1996;2:753–759.
41. Wang DG, Fan JB, Siao CJ, et al. Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 1998;280:1077–1082.
42. Winzeler EA, Richards DR, Conway AR, et al. Direct allelic variation scanning of the yeast genome. Science 1998;281:1194–1197.
43. Chhabra RS, Huff JE, Schwetz BS, Selkirk J. An overview of prechronic and chronic toxicity carcinogenicity experimental-study designs and criteria used by the National Toxicology Program. Environ Health Perspect 1990;86:313–321.
44. Ermolaeva O, Rastogi M, Pruitt KD, et al. Data management and analysis for gene expression arrays. Nat Genet 1998;20:19–23.
45. http://www.nhgri.nih.gov/DIR/LCG/15K/HTML/dbase.html
46. Samson M, Libert F, Doranz BJ, et al. Resistance to HIV-1 infection in Caucasian individuals bearing mutant alleles of the CCR-5 chemokine receptor gene. Nature 1996;382:722–725.
47. Bell DA, Taylor JA, Paulson DF, Robertson CN, Mohler JL, Lucier GW. Genetic risk and carcinogen exposure—A common inherited defect of the carcinogen-metabolism gene glutathione-S-transferase M1 (Gstm1) that increases susceptibility to bladder cancer. J Natl Cancer Inst 1993;85:1159–1164.
48. http://www.niehs.nih.gov/envgenom/home.html

Expression profiling in toxicology — potentials and limitations

Sandra Steiner *, N. Leigh Anderson

Large Scale Biology Corporation, 9620 Medical Center Drive, Rockville, MD 20850-3338, USA

Abstract

Recent progress in genomics and proteomics technologies has created a unique opportunity to significantly impact the pharmaceutical drug development processes. The perception that cells and whole organisms express specific inducible responses to stimuli such as drug treatment implies that unique expression patterns, molecular fingerprints, indicative of a drug's efficacy and potential toxicity are accessible. The integration into state-of-the-art toxicology of assays allowing one to profile treatment-related changes in gene expression patterns promises new insights into mechanisms of drug action and toxicity. The benefits will be improved lead selection, and optimized monitoring of drug efficacy and safety in pre-clinical and clinical studies based on biologically relevant tissue and surrogate markers. © 2000 Elsevier Science Ireland Ltd. All rights reserved.

Keywords: Proteomics; Genomics; Toxicology

1. Introduction

The majority of drugs act by binding to protein targets, most to known proteins representing enzymes, receptors and channels, resulting in effects such as enzyme inhibition and impairment of signal transduction. The treatment-induced perturbations provoke feedback reactions aiming to compensate for the stimulus, which almost always are associated with signals to the nucleus, resulting in altered gene expression. Such gene expression regulations account for both the pharmacological action and the toxicity of a drug and can be visualized by either global mRNA or global protein expression profiling. Hence, for each individual drug, a characteristic gene regulation pattern, its molecular fingerprint, exists which bears valuable information on its mode of action and its mechanism of toxicity.

Gene expression is a multistep process that results in an active protein (Fig. 1). There exist numerous regulation systems that exert control at and after the transcription and the translation step. Genomics, by definition, encompasses the quantitative analysis of transcripts at the mRNA level, while the aim of proteomics is to quantify gene expression further down-stream, creating a snapshot of gene regulation closer to ultimate cell function control.

---

* Corresponding author. Tel.: + 1-301-4245989; fax: + 1-301-7624892.

E-mail address: steiner@lsbc.com (S. Steiner)

2. Global mRNA profiling

Expression data at the mRNA level can be produced using a set of different technologies such as DNA microarrays, reverse transcript imaging, amplified fragment length polymorphism (AFLP), serial analysis of gene expression (SAGE) and others. Currently, DNA microarrays are very popular and promise a great potential. On a typical array, each gene of interest is represented either by a long DNA fragment (200–2400 bp) typically generated by polymerase chain reaction (PCR) and spotted on a suitable substrate using robotics (Schena et al., 1995; Shalon et al., 1996) or by several short oligonucleotides (20–30 bp) synthesized directly onto a solid support using photolabile nucleotide chemistry (Fodor et al., 1991; Chee et al., 1996). From control and treated tissues, total RNA or mRNA is isolated and reverse transcribed in the presence of radioactive or fluorescent labeled nucleotides, and the labeled probes are then hybridized to the arrays. The intensity of the array signal is measured for each gene transcript by either autoradiography or laser scanning confocal microscopy. The ratio between the signals of control and treated samples reflect the relative drug-induced change in transcript abundance.

3. Global protein profiling

Global quantitative expression analysis at the protein level is currently restricted to the use of two-dimensional gel electrophoresis. This technique combines separation of tissue proteins by isoelectric focusing in the first dimension and by sodium dodecyl sulfate slab gel electrophoresis-based molecular weight separation on the second, orthogonal dimension (Anderson et al., 1991). The product is a rectangular pattern of protein spots that are typically revealed by Coomassie Blue, silver or fluorescent staining (Fig. 2). Protein spots are identified by mass spectrometry following generation of peptide mass fingerprints (Mann et al., 1993) and sequence tags (Wilkins et al., 1996). Similar to the mRNA approach, the ratio between the optical density of spots from control and treated samples are compared to search for treatment-related changes.

4. Expression data analysis

Bioinformatics forms a key element required to organize, analyze and store expression data from either source, the mRNA or the protein level. The overall objective, once a mass of high-quality
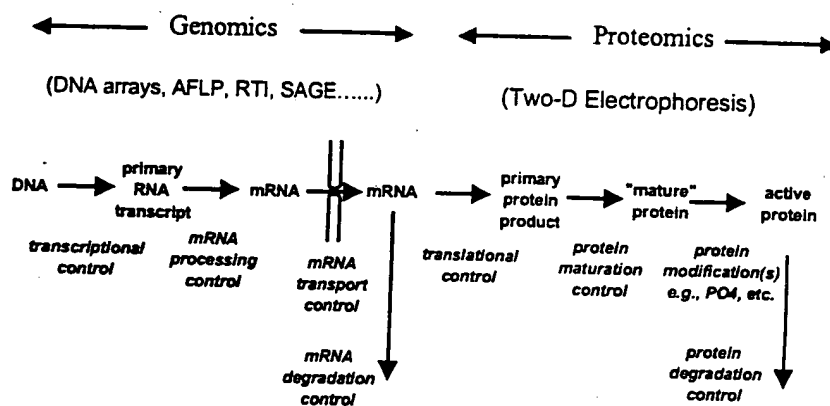

Fig. 1. Production of an active protein is a multistep process in which numerous regulation systems exert control at various stages of expression. Molecular fingerprints of drugs can be visualized through expression profiling at the mRNA level (genomics) using a variety of technologies and at the protein level (proteomics) using two-dimensional gel electrophoresis.
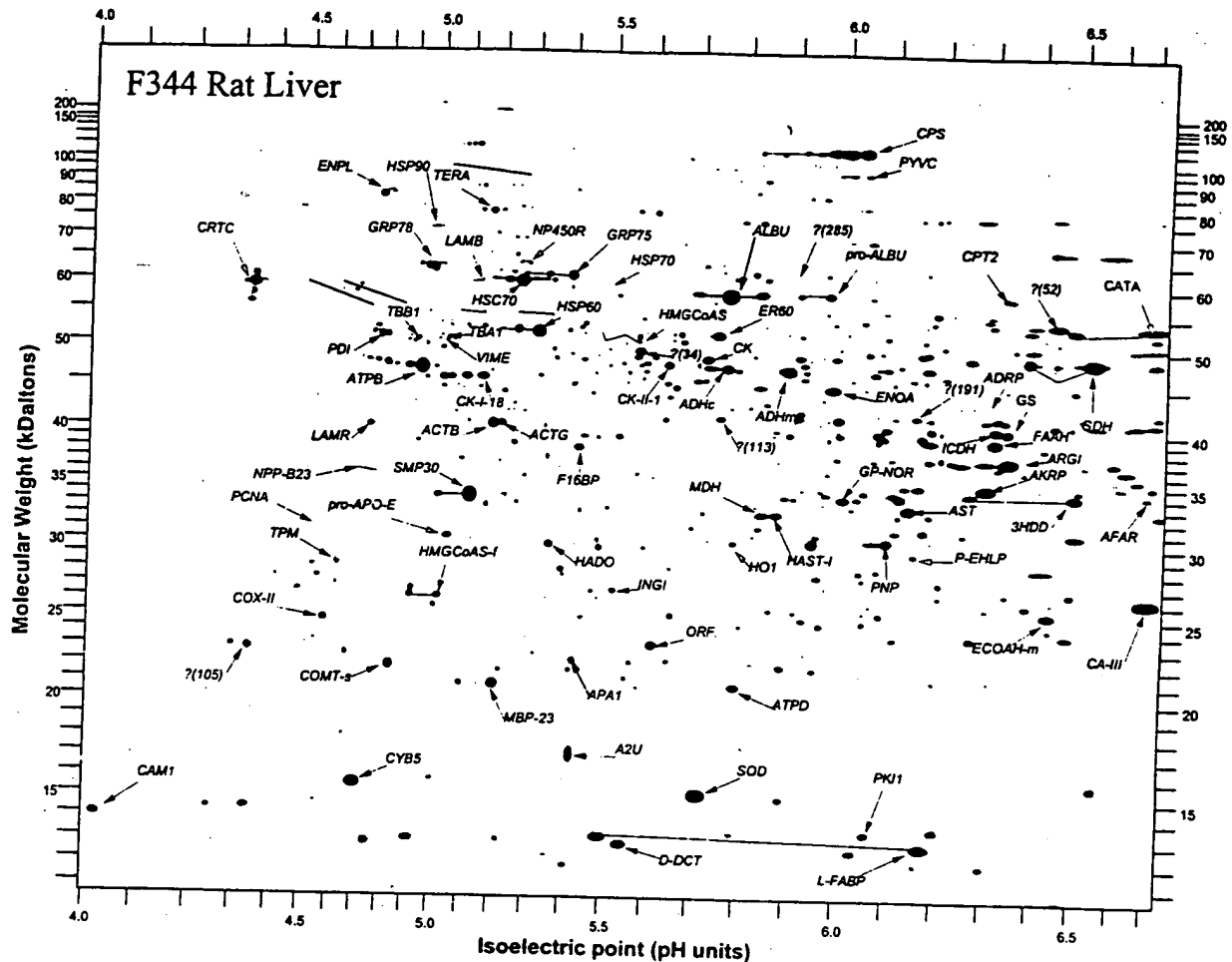
Fig. 2. Computerized representation of a Coomassie Blue stained two-dimensional gel electrophoresis pattern of Fischer F344 rat liver homogenate.

quantitative expression data has been collected, is to visualize complex patterns of gene expression changes, to detect pathways and sets of genes tightly correlated with treatment efficacy and toxicity, and to compare the effects of different sets of treatment (Anderson et al., 1996). As the drug effect database is growing, one may detect similarities and differences between the molecular fingerprints produced by various drugs, information that may be crucial to make a decision whether to refocus or extend the therapeutic spectrum of a drug candidate.

5. Comparison of global mRNA and protein expression profiling

There are several synergies and overlaps of data obtained by mRNA and protein expression analysis. Low abundant transcripts may not be easily quantified at the protein level using standard two-dimensional gel electrophoresis analysis and their detection may require prefractionation of samples. The expression of such genes may be preferably quantified at the mRNA level using techniques allowing PCR-mediated target amplifi-

cation. Tissue biopsy samples typically yield good quality of both mRNA and proteins; however, the quality of mRNA isolated from body fluids is often poor due to the faster degradation of mRNA when compared with proteins. RNA samples from body fluids such as serum or urine are often not very 'meaningful', and secreted proteins are likely more reliable surrogate markers for treatment efficacy and safety. Detection of post-translational modifications, events often related to function or nonfunction of a protein, is restricted to protein expression analysis and rarely can be predicted by mRNA profiling. Information on subcellular localization and translocation of proteins has to be acquired at the level of the protein in combination with sample prefractionation procedures. The growing evidence of a poor correlation between mRNA and protein abundance (Anderson and Seilhamer, 1997) further suggests that the two approaches, mRNA and protein profiling, are complementary and should be applied in parallel.

6. Expression profiling and drug development

Understanding the mechanisms of action and toxicity, and being able to monitor treatment efficacy and safety during trials is crucial for the successful development of a drug. Mechanistic insights are essential for the interpretation of drug effects and enhance the chances of recognizing potential species specificities contributing to an improved risk profile in humans (Richardson et al., 1993; Steiner et al.. 1996b; Aicher et al., 1998). The value of expression profiling further increases when links between treatment-induced expression profiles and specific pharmacological and toxic endpoints are established (Anderson et al., 1991, 1995, 1996; Steiner et al. 1996a). Changes in gene expression are known to precede the manifestation of morphological alterations, giving expression profiling a great potential for early compound screening, enabling one to select drug candidates with wide therapeutic windows reflected by molecular fingerprints indicative of high pharmacological potency and low toxicity (Arce et al., 1998). In later phases of drug devel-

opment, surrogate markers of treatment efficacy and toxicity can be applied to optimize the monitoring of pre-clinical and clinical studies (Doherty et al., 1998).

7. Perspectives

The basic methodology of safety evaluation has changed little during the past decades. Toxicity in laboratory animals has been evaluated primarily by using hematological, clinical chemistry and histological parameters as indicators of organ damage. The rapid progress in genomics and proteomics technologies creates a unique opportunity to dramatically improve the predictive power of safety assessment and to accelerate the drug development process. Application of gene and protein expression profiling promises to improve lead selection, resulting in the development of drug candidates with higher efficacy and lower toxicity. The identification of biologically relevant surrogate markers correlated with treatment efficacy and safety bears a great potential to optimize the monitoring of pre-clinical and clinical trails.

References

Aicher. L., Wahl. D.. Arce, A.. Grenet. O.. Steiner. S., 1998. New insights into cyclosporine A nephrotoxicity by proteome analysis. Electrophoresis 19, 1998-2003.

Anderson. N.L.. Seilhamer. J., 1997. A comparison of selected mRNA and protein abundances in human liver. Electrophoresis 18. 533-537.

Anderson. N.L.. Esquer-Blasco. R., Hofmann. J.P., Anderson, N.G.. 1991. A two-dimensional gel database of rat liver proteins useful in gene regulation and drug effects studies. Electrophoresis 12. 907-930.

Anderson. L.. Steele. V.K.. Kelloff. G.J., Sharma. S., 1995. Effects of oltipraz and related chemoprevention compounds on gene expression in rat liver. J. Cell. Biochem. Suppl. 22. 108-116.

Anderson. N.L., Esquer-Blasco, R.. Richardson. F.. Foxworthy. P.. Eacho. P., 1996. The effects of peroxisome proliferators on protein abundances in mouse liver. Toxicol. Appl. Pharmacol. 137. 75-89.

Arce. A.. Aicher. L., Wahl, D.. Esquer-Blasco, R., Anderson, N.L.. Cordier. A.. Steiner. S., 1998. Changes in the liver proteome of female Wistar rats treated with the hypoglycemic agent SDZ PGU 693. Life Sci. 63. 2243-2250.

Chee, M., Yang, R., Hubbell, E., Berno, A., Huang, X.C., Stern, D., Winkler, J., Lockhart, D.J., Morris, M.S., Fodor, S.P., 1996. Accessing genetic information with high-density DNA arrays. Science 274, 610–614.

Doherty, N.S., Littman, B.H., Reilly, K., Swindell, A.C., Buss, J., Anderson, N.L., 1998. Analysis of changes in acute-phase plasma proteins in an acute inflammatory response and in rheumatoid arthritis using two-dimensional gel electrophoresis. Electrophoresis 19, 355–363.

Fodor, S.P., Read, J.L., Pirrung, M.C., Stryer, L., Lu, A.T., Solas, D., 1991. Light-directed, spatially addressable parallel chemical synthesis. Science 251, 767–773.

Mann, M., Hojrup, P., Roepsdorff, P., 1993. Use of mass spectrometric molecular weight information to identify proteins in sequence databases. Biol. Mass Spectrom. 22, 338–345.

Richardson, F.C., Strom, S.C., Copple, D.M., Bendele, R.A., Probst, G.S., Anderson, N.L., 1993. Comparisons of protein changes in human and rodent hepatocytes induced by the rat-specific carcinogen, methapyrilene. Electrophoresis 14, 157–161.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O., 1995. Quantitative monitoring of gene expresssion patterns with a complementary DNA microarray. Science 251, 467–470.

Shalon, D., Smith, S.J., Brown, P.O., 1996. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 6, 639–645.

Steiner, S., Wahl, D., Mangold, B.L.K., Robison, R., Rayrnackers, J., Meheus, L., Anderson, N.L., Cordier, A., 1996a. Induction of the adipose differentiation-related protein in liver of etomoxir treated rats. Biochem. Biophys. Res. Commun. 218, 777–782.

Steiner, S., Aicher, L., Raymackers, J., Meheus, L., Esquer-Blasco, R., Anderson, L., Cordier, A., 1996b. Cyclosporine A mediated decrease in the rat renal calcium binding protein calbindin-D 28 kDa. Biochem. Pharmacol. 51, 253–258.

Wilkins, M.R., Gasteiger, E., Sanchez, J.C., Appel, R.D., Hochstrasser, D.F., 1996. Protein identification with sequence tags. Curr. Biol. 6, 1543–1544.

Application of DNA Arrays to Toxicology

John C. Rockett and David J. Dix

Reproductive Toxicology Division, National Health and Environmental Effects Research Laboratory, U.S. Environmental Protection Agency, Research Triangle Park, North Carolina, USA

DNA array technology makes it possible to rapidly genotype individuals or quantify the expression of thousands of genes on a single filter or glass slide, and holds enormous potential in toxicologic applications. This potential led to a U.S. Environmental Protection Agency-sponsored workshop titled "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. In addition to providing state-of-the-art information on the application of DNA or gene microarrays, the workshop catalyzed the formation of several collaborations, committees, and user's groups throughout the Research Triangle Park area and beyond. Potential application of microarrays to toxicologic research and risk assessment include genome-wide expression analyses to identify gene-expression networks and toxicant-specific signatures that can be used to define mode of action, for exposure assessment, and for environmental monitoring. Arrays may also prove useful for monitoring genetic variability and its relationship to toxicant susceptibility in human populations. *Key words:* DNA arrays, gene arrays, microarrays, toxicology. *Environ Health Perspect* 107:681–685 (1999). [Online 6 July 1999]
http://ehpnet1.niehs.nih.gov/docs/1999/107p681-685rockett/abstract.html

Decoding the genetic blueprint is a dream that offers manifold returns in terms of understanding how organisms develop and function in an often hostile environment. With the rapid advances in molecular biology over the last 30 years, the dream has come a step closer to reality. Molecular biologists now have the ability to elucidate the composition of any genome. Indeed, almost 20 genomes have already been sequenced and more than 60 are currently under way. Foremost among these is the Human Genome Mapping Project. However, the genomes of a number of commonly used laboratory species are also under intensive investigation, including yeast, *Arabidopsis*, maize, rice, zebra fish, mouse, rat, and dog. It is widely expected that the completion of such programs will facilitate the development of many powerful new techniques and approaches to diagnosing and treating genetically and environmentally induced diseases which afflict mankind. However, the vast amount of data being generated by genome mapping will require new high-throughput technologies to investigate the function of the millions of new genes that are being reported. Among the most widely heralded of the new functional genomics technologies are DNA arrays, which represent perhaps the most anticipated new molecular biology technique since polymerase chain reaction (PCR).

Arrays enable the study of literally thousands of genes in a single experiment. The potential importance of arrays is enormous and has been highlighted by the recent publication of an entire *Nature Genetics* supplement dedicated to the technology (*1*). Despite this huge surge of interest, DNA arrays are still little used and largely unproven, as demonstrated by the high ratio of review and press articles to actual data papers. Even so, the potential they offer has driven venture capitalists into a frenzy of investment and many new companies are springing up to claim a share of this rapidly developing market.

The U.S. Environmental Protection Agency (EPA) is interested in applying DNA array technology to ongoing toxicologic studies. To learn more about the current state of the technology, the Reproductive Toxicology Division (RTD) of the National Health and Environmental Effects Research Laboratory (NHEERL; Research Triangle Park, NC) hosted a workshop on "Application of Microarrays to Toxicology" on 7–8 January 1999 in Research Triangle Park, North Carolina. The workshop was organized by David Dix, Robert Kavlock, and John Rockett of the RTD/NHEERL. Twenty-two intramural and extramural scientists from government, academia, and industry shared information, data, and opinions on the current and future applications for this exciting new technology. The workshop had more than 150 attendees, including researchers, students, and administrators from the EPA, the National Institute of Environmental Health Sciences (NIEHS), and a number of other establishments from Research Triangle Park and beyond. Presentations ranged from the technology behind array production through the sharing of actual experimental data and projections on the future importance and applications of arrays. The information contained in the workshop presentations should provide aid and insight into arrays in general and their application to toxicology in particular.

Array Elements

In the context of molecular biology, the word "array" is normally used to refer to a series of DNA or protein elements firmly attached in a regular pattern to some kind of supportive medium. DNA array is often used interchangeably with gene array or microarray. Although not formally defined, microarray is generally used to describe the higher density arrays typically printed on glass chips. The DNA elements that make up DNA arrays can be oligonucleotides, partial gene sequences, or full-length cDNAs. Companies offering pre-made arrays that contain less than full-length clones normally use regions of the genes which are specific to that gene to prevent false positives arising through cross-hybridization. Sequence verification of cDNA clone identity is necessary because of errors in identifying specific clones from cDNA libraries and databases. Premade DNA arrays printed on membranes are currently or imminently available for human, mouse, and rat. In most cases they contain DNA sequences representing several thousand different sequence clusters or genes as delineated through the National Center for Biotechnology Information UniGene Project (*2*). Many of these different UniGene clusters (putative genes) are represented only by expressed sequence tags (ESTs).

Array Printing

Arrays are typically printed on one of two types of support matrix. Nylon membranes are used by most off-the-shelf array providers such as Clontech Laboratories, Inc. (Palo Alto, CA), Genome Systems, Inc. (St. Louis, MO), and Research Genetics, Inc. (Huntsville, AL). Microarrays such as those produced by Affymetrix, Inc. (Santa Clara, CA), Incyte Pharmaceuticals, Inc. (Palo Alto, CA), and many do-it-yourself (DIY) arraying groups use glass wafers or slides. Although standard microscope slides may be used, they must be preprepared to facilitate sticking of the DNA to the glass. Several different

coatings have been successfully used, including silane and lysine. The coating of slides can easily be carried out in the laboratory, but many prefer the convenience of precoated slides available from suppliers.

Once the support matrix has been prepared, the DNA elements can be applied by several methods. Affymetrix, Inc., has developed a unique photolithographic technology for attaching oligonucleotides to glass wafers. More commonly, DNA is applied by either noncontact or contact printing. Noncontact printers can use thermal, solenoid, or piezoelectric technology to spray aliquots of solution onto the support matrix and may be used to produce slide or membrane-based arrays. Cartesian Technologies, Inc. (Irvine, CA) has developed nQUAD technology for use in its PixSys printers. The system couples a syringe pump with the microsolenoid valve, a combination that provides rapid quantitative dispensing of nanoliter volumes (down to 4.2 nL) over a variable volume range. A different approach to noncontact printing uses a solid pin and ring combination (Genetic MicroSystems, Inc., Woburn, MA). This system (Figure 1) allows a broader range of sample, including cell suspensions and particulates, because the printing head cannot be blocked up in the same way as a spray nozzle. Fluid transfer is controlled in this system primarily by the pin dimensions and the force of deposition, although the nature of the support matrix and the sample will also affect transfer to some degree.

In contact printing, the pin head is dipped in the sample and then touched to the support matrix to deposit a small aliquot. Split pins were one of the first contact-printing devices to be reported and are the suggested format for DIY arrayers, as described by Brown (3). Split pins are small metal pins with a precise groove cut vertically in the middle of the pin tip. In this system, 1–48 split pins are positioned in the pin-head. The split pins work by simple capillary action, not unlike a fountain pen—when the pin heads are dipped in the sample, liquid is drawn into the pin groove. A small (fixed) volume is then deposited each time the split pins are gently touched to the support matrix. Sample (100–500 pL depending on a variety of parameters) can be deposited on multiple slides before refilling is required, and array densities of > 2,500 spots/cm² may be produced. The deposit volume depends on the split size, sample fluidity, and the speed of printing. Split pins are relatively simple to produce and can be made in-house if a suitable machine shop is available. Alternatively, they can be obtained directly from companies such as TeleChem International, Inc. (Sunnyvale, CA).

Irrespective of their source, printers should be run through a preprint sequence prior to producing the actual experimental

arrays; the first 100 or so spots of a new run tend to be somewhat variable. Factors effecting spot reproducibility include slide treatment homogeneity, sample differences, and instrument errors. Other factors that come into play include clean ejection of the drop and clogging (nQUAD printing) and mechanical variations and long-term alteration in print-head surface of solid and split pins. However, with careful preparation it is possible to get a coefficient of variance for spot reproducibility below 10%.

One potential printing problem is sample carryover. Repeated washing, blotting, and drying (vacuum) of print pins between samples is normally effective at reducing sample carryover to negligible amounts. Printing should also be carried out in a controlled environment. Humidified chambers are available in which to place printers. These help prevent dust contamination and produce a uniform drying rate, which is important in determining spot size, quality, and reproducibility.

In summary, although several printing technologies are available, none are particularly outstanding and the bottom line is that they are still in a relatively early stage of evolution.

Array Hybridization

The hybridization protocol is, practically speaking, relatively straightforward and those with previous experience in blotting should have little difficulty. Array hybridizations are, in essence, reverse Southern/Northern blots—instead of applying a labeled probe to the target population of DNA/RNA, the labeled population is applied to the probe(s). With membrane-based arrays, the control and treated mRNA populations are normally converted to cDNA and labeled with isotope (e.g., 33P) in the process. These labeled populations are then hybridized independently to parallel or serial arrays and the hybridization signal is detected with a phosporimager. A less commonly used alternative to radioactive probes is enzymatic detection. The probe may be biotinylated, haptenylated, or have alkaline phosphatase/horseradish peroxidase attached. Hybridization is detected by enzymatic reaction yielding a color reaction (4). Differences in hybridization signals can be detected by eye or, more accurately, with the help of digital imaging and commercially available software. The labeling of the test populations for slide-based microarrays uses a slightly different approach. The probe typically consists of two samples of polyA+ RNA (usually from a treated and a control population) that are converted to cDNA; in the process each is labeled with a different fluor. The independently labeled probes are then mixed together and hybridized to a single microarray slide and the resulting combined fluorescent signal is scanned. After
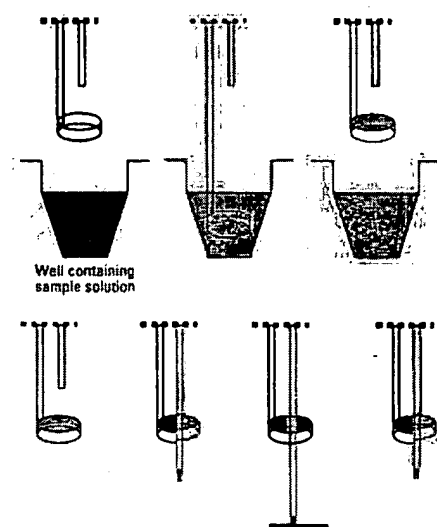

Figure 1. Genetic Microsystems (Woburn, MA) pin ring system for printing arrays. The pin ring combination consists of a circular open ring oriented parallel to the sample solution, with a vertical pin centered over the ring. When the ring is dipped into a solution and lifted, it withdraws an aliquot of sample held by surface tension. To spot the sample, the pin is driven down through the ring and a portion of the solution is transferred to the bottom of the pin. The pin continues to move downward until the pendant drop of solution makes contact with the underlying surface. The pin is then lifted, and gravity and surface tension cause deposition of the spot onto the array. Figure from Flowers et al. (14), with permission from Genetic Microsystems.

normalization, it is possible to determine the ratio of fluorescent signals from a single hybridization of a slide-based microarray.

cDNA derived from control and treated populations of RNA is most commonly hybridized to arrays, although subtractive hybridization or differential display reactions may also be used. Fluorophore- or radiolabeled nucleotides are directly incorporated into the cDNA in the process of converting RNA to cDNA. Alternatively, 5' end-labeled primers may be used for cDNA synthesis. These are labeled with a fluorophore for direct visualization of the hybridized array. Alternatively, biotin or a hapten may be attached to the primer, in which case fluor-labeled streptavidin or antibody must be applied before a signal can be generated. The most commonly used fluorophores at present are cyanine (Cy)3 and Cy5 (Amersham Pharmacia Biotech AB, Uppsala, Sweden). However, the relative expense of these fluorescent conjugates has driven a search for cheaper alternatives. Fluorescein, rhodamine, and Texas red have all been used, and companies such as Molecular Probes, Inc. (Eugene, OR) are developing a series of labeled nucleotides with a wide range of excitation and emission spectra which may prove to function as well as the Cy dyes.

Analysis of DNA Microarrays

Membrane-based arrays are normally analyzed on film or with a phosphorimager, whereas chip-based arrays require more specialized scanning devices. These can be divided into three main groups: the charge-coupled device camera systems, the nonconfocal laser scanners, and the confocal laser scanners. The advantages and disadvantages of each system are listed in Table 1.

Because a typical spot on a microarray can contain > 10^8 molecules, it is clear that a large variation in signal strength may occur. Current scanners cannot work across this many orders of magnitude (4 or 5 is more typical). However, the scanning parameters can normally be adjusted to collect more or less signal, such that two or three scans of the same array should permit the detection of rare and abundant genes.

When a microarray is scanned, the fluorescent images are captured by software normally included with the scanner. Several commercial suppliers provide additional software for quantifying array images, but the software tools are constantly evolving to meet the developing needs of researchers, and it is prudent to define one's own needs and clarify the exact capabilities of the software before its purchase. Issues that should be considered include the following:

- Can the software locate offset spots?
- Can it quantitate across irregular hybridization signals?
- Can the arrayed genes be programmed in for easy identification and location?
- Can the software connect via the Internet to databases containing further information on the gene(s) of interest?

One of the key issues raised at the workshop was the sensitivity of microarray technology. Experiments by General Scanning, Inc. (Watertown, MA), have shown that by using the Cy dyes and their scanner, signal can be detected down to levels of < 1 fluor molecule per square micrometer, which translates to detecting a rare message at approximately one copy per cell or less.

Array Applications

Although arrays are an emerging technology certain to undergo improvement and alteration, they have already been applied usefully to a number of model systems. Arrays are at their most powerful when they contain the entire genome of the species they are being used to study. For this reason, they have strong support among researchers utilizing yeast and *Caenorhabditis elegans* (5). The genomes of both of these species have been sequenced and, in the case of yeast, deposited onto arrays for examination of gene expression (6,7). With both of these species, it is relatively easy to perturb individual gene expression. Indeed, *C.*

elegans knockouts can be made simply by soaking the worms in an antisense solution of the gene to be knocked out.

By a process of systematic gene disruption, it is now possible to examine the cause and effect relationships between different genes in these simple organisms. This kind of approach should help elucidate biochemical pathways and genetic control processes, deconvolute polygenic interactions, and define the architecture of the cellular network. A simple case study of how this can be achieved was presented by Butow [University of Texas Southwestern Medical Center, Dallas, TX (Figure 2)]. Although it is the phenotypic result of a single gene knockout that is being examined, the effect of such perturbation will almost always be polygenic. Polygenic interactions will become increasingly important as researchers begin to move away from single gene systems when examining the nature of toxicologic responses to external stimuli. This is especially important in toxicology because the phenotype produced by a given environmental insult is never the result of the action of a single gene; rather, it is a complex interaction of one or multiple cellular pathways. Phenomena such as quantitative trait (the continuous variation of phenotype), epistasis (the effect of alleles of one or more genes on the expression of other genes), and penetrance (proportion of individuals of a given genotype that display a particular phenotype) will become increasingly evident and important as toxicologists push toward the ultimate goal of matching the responses of individuals to different environmental stimuli.

Analysis of the transcriptome (the expression level of all the genes in a given cell population) was a use of arrays addressed by several speakers. Unfortunately, current gene nomenclature is often confusing in that single genes are allocated multiple names (usually as a result of independent discovery by different laboratories), and there was a call for standardization of gene nomenclature. Nevertheless, once a transcriptome has been assembled it can then be transferred onto arrays and used to screen any chosen system. The EPA MicroArray Consortium (EPAMAC) is assembling testes

transcriptomes for human, rat, and mouse. In a slightly different approach, Nuwaysir et al. (8) describes how the NIEHS assembled what is effectively a "toxicological transcriptome"—a library of human and mouse genes that have previously been proven or implicated in responses to toxicologic insults. Clontech Laboratories, Inc. (Palo Alto, CA), has begun a similar process by developing stress/toxicology filter arrays of rat, mouse, and human genes. Thus, rather than being tissue or cell specific, these stress/toxicology arrays can be used across a variety of model systems to look for alterations in the expression of toxicologically important genes and define the new field of toxicogenomics. The potential to identify toxicant families based on tissue- or cell-specific gene expression could revolutionize drug testing. These molecular signatures or fingerprints could not only point to the possible toxicity/carcinogenicity of newly discovered compounds (Figure 3), but also aid in elucidating their mechanism of action through identification of gene expression networks. By extension, such signatures could provide easily identifiable biomarkers to assess the degree, time, and nature of exposure.

DNA arrays are primarily a tool for examining differential gene expression in a given model. In this context they are referred to as closed systems because they lack the ability of other differential expression technologies, e.g., differential display and subtractive hybridization, to detect previously unknown genes not present on the array. This would appear to limit the power of DNA arrays to the imaginations and preconceptions of the researcher in selecting genes previously characterized and thought to be involved in the model system. However, the various genome sequencing projects have created a new category of sequence—the EST—that has partially mollified this deficiency. ESTs are cDNAs expressed in a given tissue that, although they may share some degree of sequence similarity to previously characterized genes, have not been assigned specific genetic identity. By incorporating EST clones into an array, it is possible to monitor the expression of these unknown genes. This can enable the identification of previously uncharacterized genes that may have biologic

Table 1. Advantages and disadvantages of different microarray scanning systems.

	Nonconfocal laser scanner		
Advantages	Few moving parts	Relatively simple optics	Small depth of focus reduces artifacts
	Fast scanning of bright samples		May have high light collection efficiency
Disadvantages	Less appropriate for dim samples	Low light collection efficiency	Small depth of focus requires scanning precision
	Optical scatter can limit performance	Background artifacts not rejected	
		Resolution typically low	

CCD, charge-coupled device.
From Kawasaki (13).

significance in the model system. Filter arrays from Research Genetics and slide arrays from Incyte Pharmaceuticals both incorporate large numbers of ESTs from a variety of species.

A further use of microarrays is the identification of single nucleotide polymorphisms (SNPs). These genomic variations are abundant—they occur approximately every 1 kb or so—and are the basis of restriction fragment length polymorphism analysis used in forensic analysis. Affymetrix, Inc., designed chips that contain multiple repeats of the same gene sequence. Each position is present with all four possible bases. After the hybridization of the sample, the degree of hybridization to the different sequences can be measured and the exact sequence of the target gene deduced. SNPs are thought to be of vital importance in drug metabolism and toxicology. For example, single base differences in the regulatory region or active site of some genes can account for huge differences in the activity of that gene. Such SNPs are thought to explain why some people are able to metabolize certain xenobiotics better than others. Thus, arrays provide a further tool for the toxicologist investigating the nature of susceptible subpopulations and toxicologic response.

There are still many wrinkles to be ironed out before arrays become a standard tool for toxicologists. The main issues raised at the workshop by those with hands-on experience were the following:

• Expense: the cost of purchasing/contracting this technology is still too great for many individual laboratories.
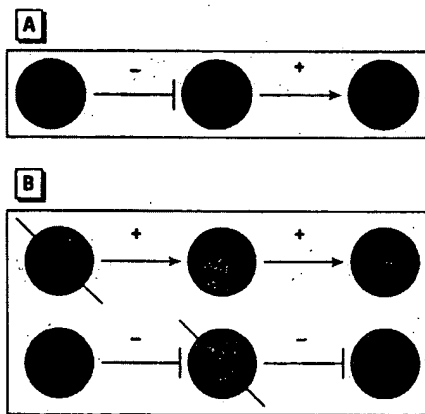
Figure 2. Potential effects of gene knockout within positively and negatively regulated gene expression networks. i_1 is limiting in wild type for expression of i_2. (A) A simple, two-component, linear regulatory network operating on gene i_2, where i_1 is a positive effector of i_2 and j_n is either a positive or negative effector of i_1. This network could be deduced by examining the consequence of (B) deleting j_n on the expression of i_1 and i_2, where the expression of i_2 would be decreased or increased depending on whether j_n was a positive or negative regulator. These and other connected components of even greater complexity could be revealed by genome-wide expression analysis. From Butow (15).

• Clones: the logistics of identifying, obtaining, and maintaining a set of nonredundant, non-contaminated, sequence-verified, species/cell/tissue/field-specific clones.
• Use of inbred strains: where whole-organism models are being used, the use of inbred strains is important to reduce the potentially confusing effects of the individual variation typically seen in outbred populations.
• Probe: the need for relatively large amounts of RNA, which limits the type of sample (e.g., biopsy) that can be used. Also, different RNA extraction methods can give different results.
• Specificity: the ability to discriminate accurately between closely related genes (e.g., the cytochrome p450 family) and splice variants.
• Quantitation: the quantitation of gene expression using gene arrays is still open to debate. One reason for this is the different incorporation of the labeling dyes. However, the main difficulty lies in knowing what to normalize against. One option is to include a large number of so-called housekeeping genes in the array. However, the expression of these genes often change depending on the tissue and the toxicant, so it is necessary to characterize the expression of these genes in the model system before utilizing them. This is clearly not a viable option when screening multiple new compounds. A second option is to include on the array genes from a nonrelated species (e.g., a plant gene on an animal array) and to spike the probe with synthetic RNA(s) complementary to the gene(s).
• Reproducibility: this is sometimes questionable, and a figure of approximately two or three repeats was used as the minimum number required to confirm initial findings.

Again, however, most people advocated the use of Northern blots or reverse transcriptase PCR to confirm findings.
• Sensitivity: concerns were voiced about the number of target molecules that must be present in a sample for them to be detected on the array.
• Efficiency: reproducible identification of 1.5- to 2-fold differences in expression was reported, although the number of genes that undergo this level of change and remain undetected is open to debate. It is important that this level of detection be ultimately achieved because it is commonly perceived that some important transcription factors and their regulators respond at such low levels. In most cases, 3- to 5-fold was the minimum change that most were happy to accept.
• Bioinformatics: perhaps the greatest concern was how to accurately interpret the data with the greatest accuracy and efficiency. The biggest headache is trying to identify networks of gene expression that are common to different treatments or doses. The amount of data from a single experiment is huge. It may be that, in the future, several groups individually equipped with specialized software algorithms for studying their favorite genes or gene systems will be able to share the same hybridized chips. Thus, arrays could usher in a new perspective on collaboration and the sharing of data.

EPAMAC

Perhaps the main reason most scientists are unable to use array technology is the high cost involved, whether buying off-the-shelf membranes, using contract printing services, or

Figure 3. Gene expression profiles—also called fingerprints or signatures—of known toxicants or toxicant families may, in the future, be used to identify the potential toxicity of new drugs, etc. In this example, the genetic signature of test compound 1 is identical to that of known peroxisome proliferators, whereas that of test compound 2 does not match any known toxicant family. Based on these results, test compound 2 would be retained for further testing and test compound 1 would be eliminated.

producing chips in-house. In view of this, researchers at the RTD/NHEERL initiated the EPAMAC. This consortium brings together scientists from the EPA and a number of extramural labs with the aim of developing microarray capability through the sharing of resources and data. EPAMAC researchers are primarily interested in the developmental and toxicologic changes seen in testicular and breast tissue, and a portion of the workshop was set aside for EPAMAC members to share their ideas on how the experimental application of microarrays could facilitate their research. One of the central areas of interest to EPAMAC members is the effect of xenobiotics on male fertility and reproductive health. Of greatest concern is the effect of exposure during critical periods of development and germ cell differentiation (9), and how this may compromise sperm counts and quality following sexual maturation (10). As well as spermatogenic tissue, there is also interest in how residual mRNA found in mature sperm (11) could be used as an indicator of previous xenobiotic effects (it is easier to obtain a semen sample than a testicular biopsy). Arrays will be used to examine and compare the effect of exposure to heat and chemicals in testicular and epididymal gene expression profiles, with the aim of establishing relationships/associations between changes in developmental landmarks and the effects on sperm count and quality. Cluster, pattern, and other analysis of such data should help identify hidden relationships between genes that may reveal potential mechanisms of action and uncover roles for genes with unknown functions.

Summary

The full impact of DNA arrays may not be seen for several years, but the interest shown at this regional workshop indicates the high level of interest that they foster. Apart from educating and advertising the various technologies in this field, this workshop brought together a number of researchers from the Research Triangle Park area who are already using DNA arrays. The interest in sharing ideas and experiences led to the initiation of a Triangle array user's group.

Array technology is still in its infancy. This means that the hardware is still improving and there is no current consensus for standard procedures, quantitation, and interpretation. Consistency in spotting and scanning arrays is not yet optimized, and this is one of the most critical requirements of any experiment. In addition, one of the dark regions of array technology—strife in the courts over who owns what portions of it—has further muddled the future and is a potential barrier toward the development of consensus procedures.

Perhaps the greatest hurdle for the application of arrays is the actual interpretation of data. No specialists in bioinformatics attended the workshop, largely because they are rare and because as yet no one seems clear on the best method of approaching data analysis and interpretation. Cross-referencing results from multiple experiments (time, dose, repeats, different animals, different species) to identify commonly expressed genes is a great challenge. In most cases, we are still a long way from understanding how the expression of gene X is related to the expression of gene Y, and ordering gene expression to delineate causal relationships.

To the ordinary scientist in the typical laboratory, however, the most immediate problem is a lack of affordable instrumentation. One can purchase premade membranes at relatively affordable prices. Although these may be useful in identifying individual genes to pursue in more detail using other methods, the numbers that would be required for even a small routine toxicology experiment prohibit this as a truly viable approach. For the toxicologist, there is a need to carry out multiple experiments—dose responses, time curves, multiple animals, and repeats. Glass-based DNA arrays are most attractive in this context because they can be prepared in large batches from the same DNA source and accommodate control and treated samples on the same chip. Another problem with current off-the-shelf arrays is that they often do not contain one or more of the particular genes a group is interested in. One alternative is to obtain and/or produce a set of custom clones and have contract printing of membranes or slides carried out by a company such as Genomic Solutions, Inc. (Ann Arbor, MI). This approach

is less expensive than laying out capital for one's own entire system, although at some point it might make economic sense to print one's own arrays.

Finally, DNA arrays are currently a team effort. They are a technology that uses a wide range of skills including engineering, statistics, molecular biology, chemistry, and bioinformatics. Because most individuals are skilled in only one or perhaps two of these areas, it appears that success with arrays may be best expected by teams of collaborators consisting of individuals having each of these skills.

Those considering array applications may be amused or goaded on by the following quote from *Fortune* magazine (12):

> Microprocessors have reshaped our economy, spawned vast fortunes and changed the way we live. Gene chips could be even bigger.

Although this comment may have been designed to excite the imagination rather than accurately reflect the truth, it is fair to say that the age of functional genomics is upon us. DNA arrays look set to be an important tool in this new age of biotechnology and will likely contribute answers to some of toxicology's most fundamental questions.

REFERENCES AND NOTES

1. The chipping forecast. Nat Genet 21(Suppl 1):3–60 (1999).
2. National Center for Biotechnology Information. The Unigene System. Available: www.ncbi.nlm.nih.gov/Schuler/UniGene [cited 22 March 1999].
3. Brown PO. The Brown Lab. Available: http://cmgm.Stanford.edu/pbrown [cited 22 March 1999].
4. Chen JJ, Wu R, Yang PC, Huang JY, Sher YP, Han MH, Kao WC, Lee PJ, Chiu TF, Chang F, et al. Profiling expression patterns and isolating differentially expressed genes by cDNA microarray system with colorimetry detection. Genomics 51:313–324 (1998).
5. Ward S. DNA Microarray Technology to Identify Genes Controlling Spermatogenesis. Available: www.mcb.arizona.edu/wardlab/microarray.html [cited 22 March 1999].
6. Marton MJ, DeRisi JL, Bennett HA, Iyer VR, Meyer MR, Roberts CJ, Stoughton R, Burchard J, Slade D, Dai H, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. Nat Med 4:1293–1301 (1998).
7. Brown PO. The Full Yeast Genome on a Chip. Available: http://cmgm.stanford.edu/pbrown/yeastchip.html [cited 22 March 1999].
8. Nuwaysir EF, Bittner M, Trent J, Barrett JC, Afshari CA. Microarrays and toxicology: the advent of toxicogenomics. Mol Carcinog 24(3):153–159 (1999).
9. Hecht NB. Molecular mechanisms of male germ cell differentiation. Bioessays 20:555–561 (1998).
10. Zacharewski TR. Timothy R. Zacharewski. Available: www.bch.msu.edu/faculty/zachar.htm [cited 22 March 1999].
11. Kramer JA, Krawetz SA. RNA in spermatozoa: implications for the alternative haploid genome. Mol Hum Reprod 3:473–478 (1997).
12. Stipp D. Gene chip breakthrough. Fortune, March 31:56–73 (1997).
13. Kawasaki E (General Scanning Instruments, Inc., Watertown, MA). Unpublished data.
14. Flowers P, Overbeck J, Mace ML Jr, Pagliughi FM, Eggers WJE, Yonkers H, Honkanen P, Montagu J, Rose SD. Development and Performance of a Novel Microarraying System Based on Surface Tension Forces. Available: http://www.geneticmicro.com/resources/html/coldspring.html [cited 22 March 1999].
15. Butow R (University of Texas Medical Center, Dallas, TX). Unpublished data.

SPEAKERS

Cindy Afshari	Abdel Elkahloun	Steve Krawetz	Jim Samet
NIEHS	Research Genetics, Inc.	Wayne State University	U.S. EPA
Linda Birnbaum	Sue Fenton	Nick Mace	Sam Ward
U.S. EPA	U.S. EPA	Genetic Microsystems, Inc.	University of Arizona
Ron Butow	Norman Hecht	Scott Mordecai	Jeff Welch
University of Texas	University of Pennsylvania	Affymetrix, Inc.	U.S. EPA
Southwestern Medical	Pat Hurban	Kevin Morgan	Reen Wu
Center	Paradigm Genetics, Inc.	Glaxo Wellcome, Inc.	University of California
Alex Chenchik	Bob Kavlock	Elaine Poplin	at Davis
Clontech Laboratories, Inc.	U.S. EPA	Research Genetics, Inc.	Tim Zacharewski
David Dix	Ernie Kawasaki	Don Rose	Michigan State University
U.S. EPA	General Scanning, Inc.	Cartesian Technologies, Inc.	

Subject: RE: [Fwd: Toxicology Chip]
Date: Mon. 3 Jul 2000 08:09:45 -0400
From: "Afshari.Cynthia" <afshari@niehs.nih.gov>
To: "Diana Hamlet-Cox" <dianahc@incyte.com>

You can see the list of clones that we have on our ToR chip at
http: manuel.niehs.nih.gov maps guest clonesrch.cfm
We selected a subset of genes (2000K) that we believed critical to tox
response and basic cellular processes and added a set of clones and ESTs to
this. We have included a set of control genes (80-) that were selected by
the NHGRI because they did not change across a large set of array
experiments. However, we have found that some of these genes change
significantly after tox treatments and are in the process of looking at the
variation of each of these 80- genes across our experiments.
Our chips are constantly changing and being updated and we hope that our
data will lead us to what the toxchip should really be.
I hope this answers your question.
Cindy Afshari

> ----------
> From: Diana Hamlet-Cox
> Sent: Monday, June 26, 2000 8:52 PM
> To: afshari@niehs.nih.gov
> Subject: [Fwd: Toxicology Chip]
>
> Dear Dr. Afshari,
>
> Since I have not yet had a response from Bill Grigg, perhaps he was not
> the right person to contact.
>
> Can you help me in this matter? I don't need to know the sequences.
> necessarily, but I would like very much to know what types of sequences
> are being used, e.g., GPCRs (more specific?), ion channels, etc.
>
> Diana Hamlet-Cox
>
> -------- Original Message --------
> Subject: Toxicology Chip
> Date: Mon, 19 Jun 2000 18:31:48 -0700
> From: Diana Hamlet-Cox <dianahc@incyte.com>
> Organization: Incyte Pharmaceuticals
> To: grigg@niehs.nih.gov
>
> Dear Colleague:
>
> I am doing literature research on the use of expressed genes as
> pharmacotoxicology markers, and found the Press Release dated February
> 29, 2000 regarding the work of the NIEHS in this area. I would like to
> know if there is a resource I can access (or you could provide?) that
> would give me a list of the 12,000 genes that are on your Human ToxChip
> Microarray. In particular, I am interested in the criteria used to
> select sequences for the ToxChip, including any control sequences
> included in the microarray.
>
> Thank you for your assistance in this request.

research focus   **REVIEWS**

Proteomics: a major new technology for the drug discovery process

Martin J. Page, Bob Amess, Christian Rohlff, Colin Stubberfield and Raj Parekh

Proteomics is a new enabling technology that is being integrated into the drug discovery process. This will facilitate the systematic analysis of proteins across any biological system or disease, forwarding new targets and information on mode of action, toxicology and surrogate markers. Proteomics is highly complementary to genomic approaches in the drug discovery process and, for the first time, offers scientists the ability to integrate information from the genome, expressed mRNAs, their respective proteins and subcellular localization. It is expected that this will lead to important new insights into disease mechanisms and improved drug discovery strategies to produce novel therapeutics.
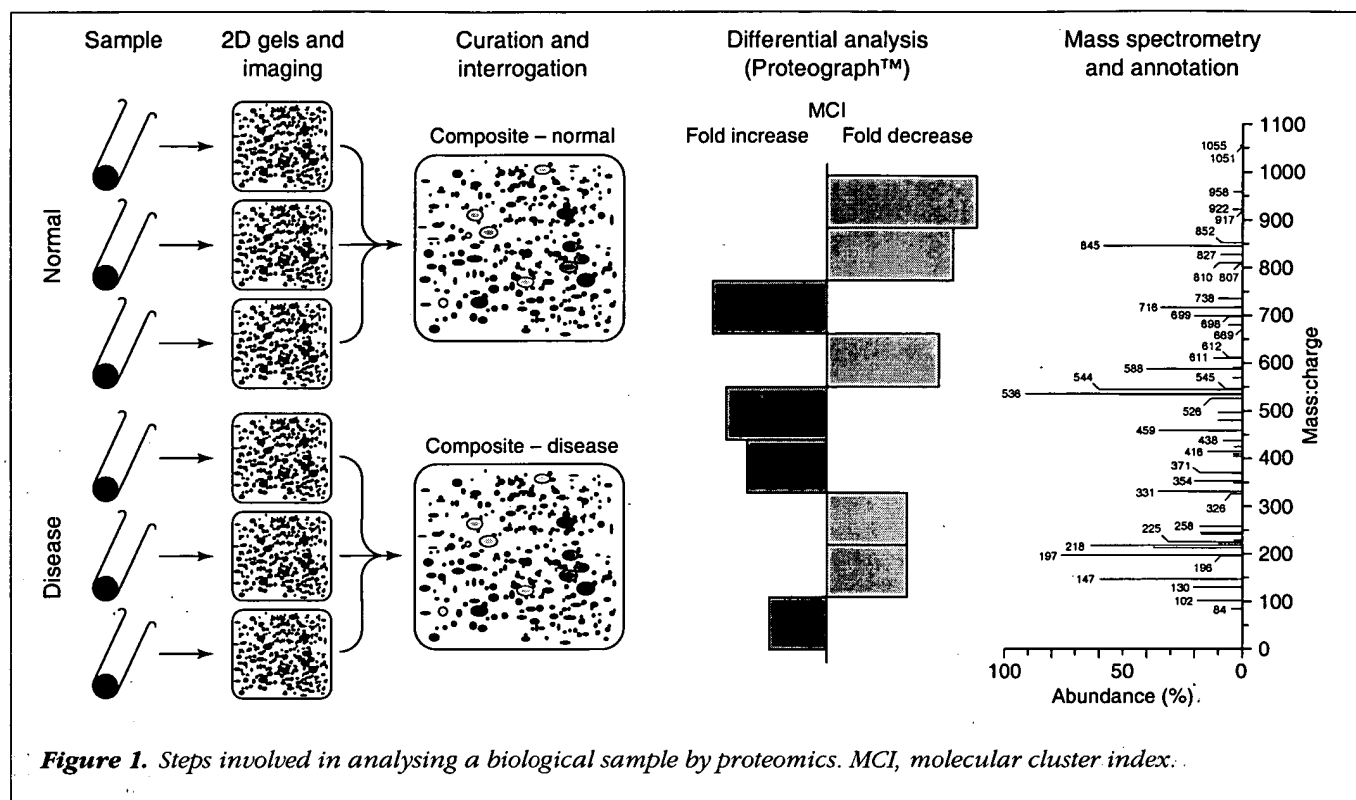
Among the major pharmaceutical and biotechnology companies, it is clearly recognized that the business of modern drug discovery is a highly competitive process. All of the many steps involved are inherently complex, and each can involve a high risk of attrition. The players in this business strive continuously to optimize and streamline the process; each seeking to gain an advantage at every step by attempting to make informed decisions at the earliest stage possible. The desired outcome is to accelerate as many key activities in the drug discovery process as possible. This should pro-duce a new generation of robust drugs that offer a high probability of success and reach the clinic and market ahead of the competition.

There has been noticeable emphasis over recent years for companies to aggressively review and refine their strategies to discover new drugs. Central to this has been the introduction and implementation of cutting-edge technologies. Most, if not all, companies have now integrated key technology platforms that incorporate genomics, mRNA expression analysis, relational databases, high-throughput robotics, combinatorial chemistry and powerful bioinformatics. Although it is still early days to quantify the real impact of these platforms in clinical and commercial terms, expectations are high, and it is widely accepted that significant benefits will be forthcoming. This is largely based on data obtained during preclinical studies where the genomic[1,2] and microarray[3,4] technologies have already proved their value.

However, there are several noteworthy outcomes that result from this. Many comments are voiced that scientists armed with these technologies are now commonly faced with data overload. Thus, in some instances, rather than facilitating the decision process, the accumulation of more complex data points, many with unknown consequences, can seem to hinder the process. Also, most drug companies have simultaneously incorporated very similar components of the new technology platforms, the consequence being that it is becoming difficult yet again to determine where a clear competitive advantage will arise. Finally, in recent years, largely as a result of the accessibility of the technologies, there has been an overwhelming emphasis placed on genomic and mRNA data rather than on protein

Martin J. Page*, Bob Am ss, Christian Rohlff, Colin Stubberfi ld and Raj Par kh, Oxford GlycoSciences, 10 The Quadrant, Abingdon Science Park, Abingdon, Oxfordshire, UK OX14 3YS. *tel: +44 1235 543277, fax: +44 1235 543283, e-mail: martin.page@ogs.co.uk

Figure 1. Steps involved in analysing a biological sample by proteomics. MCI, molecular cluster index.

analysis. It is important to remember that proteins dictate biological phenotype – whether it is normal or diseased – and are the direct targets for most drugs.

Pr t omics: new technology for the analysis of proteins

It is now timely to recognize that complementary technology in the form of high-throughput analysis of the total protein repertoire of chosen biological samples, namely proteomics, is poised to add a new and important dimension to drug discovery. In a similar fashion to genomics, which aims to profile every gene expressed in a cell, proteomics seeks to profile every protein that is expressed[5-7]. However, there is added information, since proteomics can also be used to identify the post-translational modifications of proteins[8], which can have profound effects on biological function, and their cellular localization. Importantly, proteomics is a technology that integrates the significant advances in two-dimensional (2D) electrophoretic separation of proteins, mass spectrometry and bioinformatics. With these advances it is now possible to consistently derive proteomes that are highly reproducible and suitable for interrogation using advanced bioinformatic tools.

There are many variations whereby different laboratories operate proteomics. For the purpose of this review, the process used at Oxford GlycoSciences (OGS), which uses an industrial-scale operation that is integral to its drug discovery work, will be described. The individual steps of this process, where up to 1000 2D gels can be run and analysed per week, are summarized in Fig. 1. The incoming samples are bar coded and all information relevant to the sample is logged into a Laboratory Information Management System (LIMS) database. There can be a wide range in the type of samples processed, as applicable to individual steps in the drug discovery pipeline, and these will be mentioned later. The samples are separated according to their charge (pI) in the first dimension, using isoelectric focusing, followed by size (MW) using SDS–PAGE in the second dimension. Many modifications have been made to these steps to improve handling, throughput and reproducibility. The separated proteins are then stained with fluorescent dyes which are significantly more sensitive in detection than standard silver methods and have a broader dynamic range. The image of the displayed proteins obtained is referred to as the proteome, and is digitally scanned into databases using proprietary software called ROSETTA™. The images are subsequently curated, which begins with the removal of any artefacts, cropping and the placement of pI/MW landmarks. The images from replicate images are then aligned and matched to one

another to generate a synthetic composite image. This is an important step, as the proteome is a dynamic situation, and it captures the biological variation that occurs, such that even orphan proteins are still incorporated into the analysis.

By means of illustration, Fig. 1 shows the process whereby proteomes are generated from normal and disease samples and how differentially expressed proteins are identified. The potential of this type of analysis is tremendous. For example, from a mammalian cell sample, in excess of 2000 proteins can typically be resolved within the proteome. The quality of this is shown in Fig. 2, which shows representative proteomes from three diverse biological sources: human serum, the pathogenic fungus *Candida albicans* and the human hepatoma cell line Huh7.

Us   f proteomics to identify disease specific proteins

In most cases, the drug discovery process is initiated by the identification of a novel candidate target – almost always a protein – that is believed to be instrumental in the disease process. To date, there is a variety of means whereby drug targets have been forthcoming. These include molecular, cellular and genomic approaches, mostly centred upon DNA and mRNA analysis. The gene in question is isolated, and expression and characterization of its coded protein product – i.e. the drug target – is invariably a secondary event.

With the proteomic approach, the starting point is at the other end of the 'telescope'. Here there is direct and immediate comparison of the proteomes from paired normal and disease materials. Examples of these pairs are: (1) purified epithelial cell populations derived from human breast tumours, matched to purified normal populations of human breast epithelial cells, and (2) the invading pathogenic hyphal form of *C. albicans*, matched to the non-invading yeast form of *C. albicans*. When the proteome images from each pair are aligned, the Proteograph™ software is able to rapidly identify those proteins (each referenced as having a unique molecular cluster index, or MCI) that are either unique, or those that are differentially expressed. Thus, the Proteograph output from this analysis is both qualitative and quantitative.

Proteograph analysis for a particular study can also be undertaken on any number of samples. For example, one might compare anything from a few to several hundred preparations or samples, each from a normal and disease counterpart, and have these analysed in a single Proteograph study. In this way, it is possible to assign strong statistical confidence to the data and in some instances to identify specific subpopulations within the input biological sources. This feature will become increasingly significant in the near future, and there is a clear synergy here whereby proteomics can work closely with pharmacogenomic approaches to stratify patient populations and achieve effective targeted care for the patient. Whatever the source of the materials, the net output of Proteograph analysis is immediate identification of disease specific proteins. This is shown in Fig. 3, which shows the results of a proteograph obtained by comparing untreated human hepatoma cells with cells following exposure to a clinical
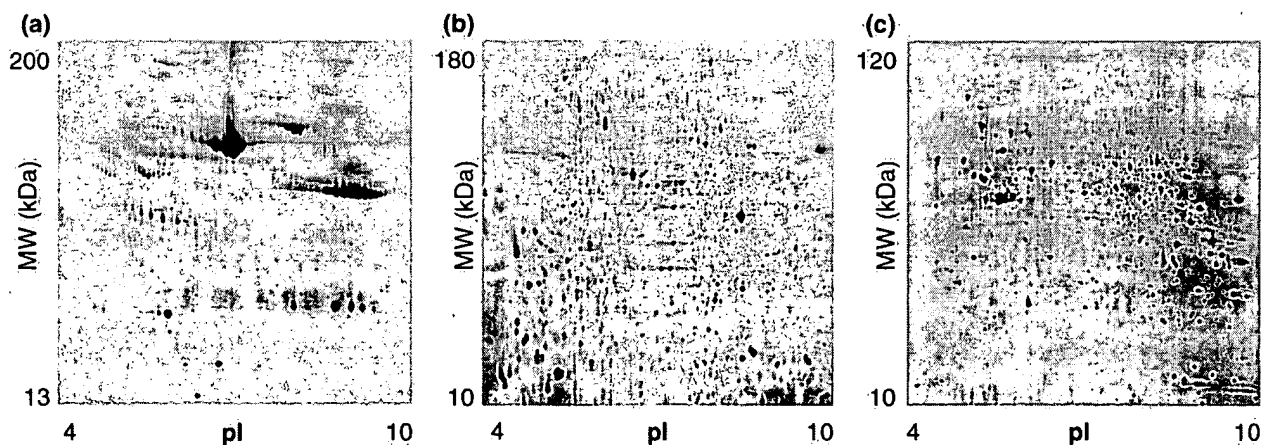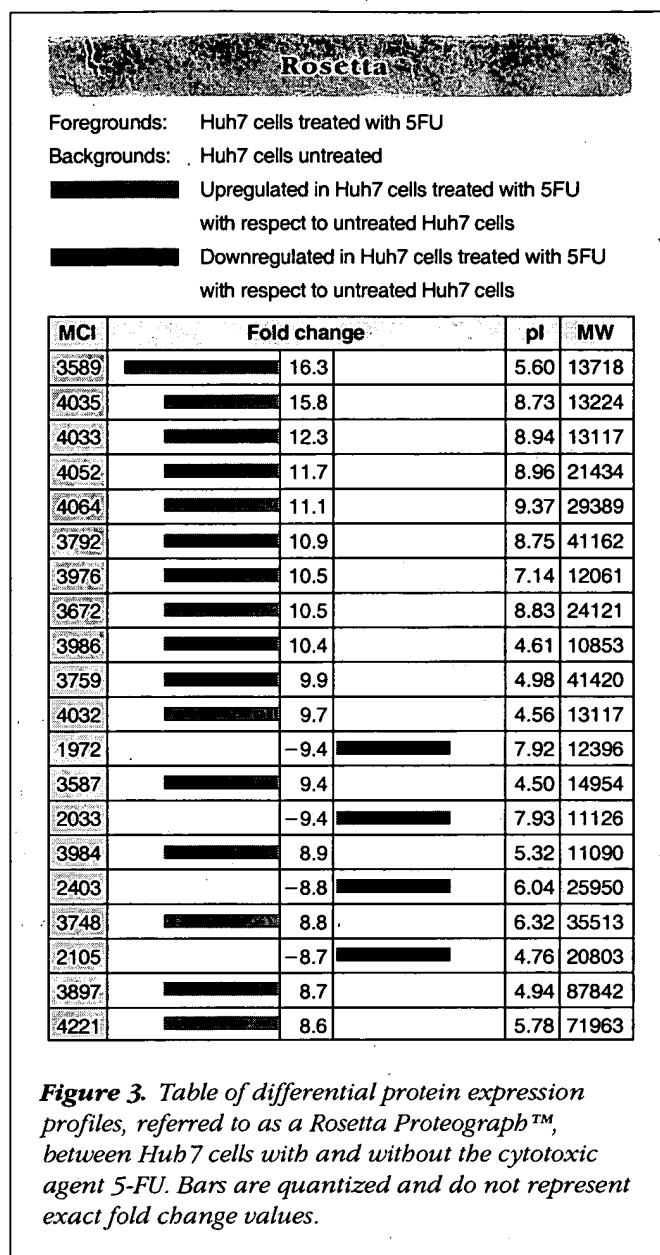

Figure 2. *Representative proteomes obtained from (a) human serum, (b) the pathogenic fungus Candida albicans and (c) the human hepatoma cell line Huh7.*

Foregrounds: Huh7 cells treated with 5FU
Backgrounds: Huh7 cells untreated

▮▮▮▮▮ Upregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

▮▮▮▮▮ Downregulated in Huh7 cells treated with 5FU with respect to untreated Huh7 cells

MCI	Fold change		pI	MW
3589	▮▮▮▮▮▮ 16.3		5.60	13718
4035	▮▮▮▮ 15.8		8.73	13224
4033	▮▮▮▮ 12.3		8.94	13117
4052	▮▮▮▮ 11.7		8.96	21434
4064	▮▮▮▮ 11.1		9.37	29389
3792	▮▮▮▮ 10.9		8.75	41162
3976	▮▮▮▮ 10.5		7.14	12061
3672	▮▮▮▮ 10.5		8.83	24121
3986	▮▮▮▮ 10.4		4.61	10853
3759	▮▮▮ 9.9		4.98	41420
4032	▮▮▮ 9.7		4.56	13117
1972	−9.4	▮▮▮▮	7.92	12396
3587	▮▮▮ 9.4		4.50	14954
2033	−9.4	▮▮▮▮	7.93	11126
3984	▮▮▮ 8.9		5.32	11090
2403	−8.8	▮▮▮▮	6.04	25950
3748	▮▮▮ 8.8		6.32	35513
2105	−8.7	▮▮▮▮	4.76	20803
3897	▮▮▮ 8.7		4.94	87842
4221	▮▮▮ 8.6		5.78	71963

Figure 3. Table of differential protein expression profiles, referred to as a Rosetta Proteograph™, between Huh7 cells with and without the cytotoxic agent 5-FU. Bars are quantized and do not represent exact fold change values.

cytotoxic agent. In this instance, only the top 20 differentially expressed MCIs are shown, but the readout would normally extend to a defined cut-off value, typically a two-fold or greater difference in expression levels, determined by the user.

In a typical analysis involving disease and normal mammalian material, in which each proteome would have ~2000 protein features each assigned an MCI, the proteograph might identify somewhere in the region of 50–300 MCIs that are unique or differentially expressed. To capitalize rapidly on these data, at OGS a high-throughput

mass spectrometry facility coupled to advanced databases to annotate these MCIs as individual proteins is applied. As these are all disease specific proteins, each could represent a novel target and/or a novel disease marker. The process becomes even more powerful when a panel of features, rather than individual features, are assigned. The relevance of this is apparent when one considers that most diseases, if not all, are multifactorial in nature and arise from polygenic changes. Rather than analysing events in isolation, the ability to examine hundreds or thousands of events simultaneously, as shown by proteomics, can offer real advantages.

Identification and assignment of candidate targets

The rapid identification and assignment of candidate targets and markers represents a huge challenge, but this has been greatly facilitated by combining the recent advances made in proteomics and analytical mass spectrometry[9]. Using automated procedures it is now possible to annotate proteins present in femtomole quantities, which would depict the low abundance class of proteins. The process of annotation is similarly aided by the quality and richness of the sequence specific databases that are currently available, both in the public domain and in the private sector (e.g. those supplied by Incyte Pharmaceuticals). In this respect, the advances in proteomics have benefited considerably from the breakthroughs achieved with genomics.

From an application perspective, cancer studies provide a good opportunity whereby proteomics can be instrumental in identifying disease specific proteins, because it is often feasible to obtain normal and diseased tissue from the same patient. For example, proteomic studies have been reported on neuroblastomas[10], human breast proteins from normal and tumour sources[11–13], lung tumours[14], colon tumours[15] and bladder tumours[16]. There are also proteomic studies reported within the cardiovascular therapeutic area, in which disease or response proteins are identified[17,18].

Genomic microarray analysis can similarly identify unique species or clusters of mRNAs that are disease specific. However, in some instances, there is a clear lack of correlation between the levels of a specific mRNA and its corresponding protein (Ref. 19, Gypi, S.P. *et al.*, submitted). This has now been noted by many investigators and reaffirms that post-transcriptional events, including protein stability, protein modification (such as phosphorylation, glycosylation, acylation and methylation) and cell localization, can constitute major regulatory steps. Proteomic analysis captures all of these steps and can therefore provide unique and valuable information independent from, or complementary to, genomic data.

Prot mics f r target validati n and signal transduc-ti n studies

The identification of disease specific proteins alone is insufficient to begin a drug screening process. It is critical to assign function and validation to these proteins by confirming they are indeed pivotal in the disease process. These studies need to encompass both gain- and loss-of-function analyses. This would determine whether the activity of a candidate target (an enzyme, for example), eliminated by molecular/cellular techniques, could reverse a disease phenotype. If this happened, then the investigator would have increased confidence that a small-molecule inhibitor against the target would also have a similar effect. The proposal of candidate drug targets is often not a difficult process, but validating them is another matter. Validation represents a major bottleneck where the wrong decision can have serious consequences[20].

Proteomics can be used to evaluate the role of a chosen target protein in signal transduction cascades directly relevant to the disease. In this manner, valuable information is forthcoming on the signalling pathways that are perturbed by a target protein and how they might be corrected by appropriate therapeutics. Techniques that are well established in one-dimensional protein studies to investigate signalling pathways, such as western blotting and immunoprecipitation, are highly suited to proteomic applications. For example, the proteomes obtained can be blotted onto membranes and probed with antibodies against the target protein or related signalling molecules[21-23]. Because proteomics can resolve >2000 proteins on a single gel, it is possible to derive important information on specific isoforms (such as glycosylated or phosphorylated variants) of signalling molecules. This will result in characterization of how they are altered in the disease process. Western immunoblotting techniques using high-affinity antibodies will typically identify proteins present at ~10 copies per cell (~1.7 fmol); this is in contrast to the best fluorescent dyes currently available that are limited to imaging proteins at 1000 or more copies per cell. The level of sensitivity derived by these applications will greatly facilitate interpretation of complex signalling pathways and contribute significantly to validation of the target under study.

Immunoprecipitation studies

Similarly, immunoprecipitation studies are another useful way to exploit the resolving power of proteomics[24,25]. In this instance, very large quantities of protein (e.g. several milligrams) can be subjected to incubation with antibodies against chosen signalling molecules. This allows high-affin-ity capture of these proteins, which can subsequently be eluted and electrophoresed on a 2D gel to provide a high-resolution proteome of a specific subset of proteins. Detection by blot analysis allows the identification of extremely small amounts of defined signalling molecules. Again, the different isoforms of even very low abundance proteins can be seen, and, very importantly, the technique allows the investigator to identify multiprotein complexes or other proteins that co-precipitate with the target protein. These coassociating proteins frequently represent signalling partners for the target protein, and their identification by mass spectrometry can lead to invaluable information on the signalling processes involved.

The depth of signal transduction analysis offered by proteomics, and the utility for target validation studies, can be extended even further by applying cell fractionation studies[26-28]. By purifying subcellular fractions, such as membrane, nuclear, organelle and cytosolic, it is possible to assign a localization to proteins of interest and to follow their trafficking in a cell. Enrichment of these fractions will also allow much higher representation of low abundance proteins on the proteome. Their detection by fluorescent dyes or immunoblot techniques will lead to the identification of proteins in the range of 1–10 copies per cell, putting the sensitivity on a par with genomic approaches.

These signal transduction analyses can be of additional value in experiments where inhibitors derived from a screening programme against the target are being evaluated for their potency and selectivity. The inhibitors can encompass small molecules, antisense nucleic acid constructs, dominant-negative proteins, or neutralizing antibodies microinjected into cells. In each case, proteome analysis can provide unique data in support of validation studies for a chosen candidate drug target.

Proteomics and drug mode-of-action studies

Once a validated target is committed to a screening regimen to identify and advance a lead molecule, it is important to confirm that the efficacy of the inhibitor is through the expected mechanism. Such mode-of-action studies are usually tackled by various cell biological and biochemical methods. Proteomics can also be usefully applied to these studies and this is illustrated below by describing data obtained with OGT719. This is a novel galactosyl derivative of the cytotoxic agent 5-fluorouracil (5-FU), which is currently being developed by OGS for the treatment of hepatocellular carcinoma and colorectal metastases localized in the liver. The premise underpinning the design and rationale of OGT719 was to derive a 5-FU prodrug capable
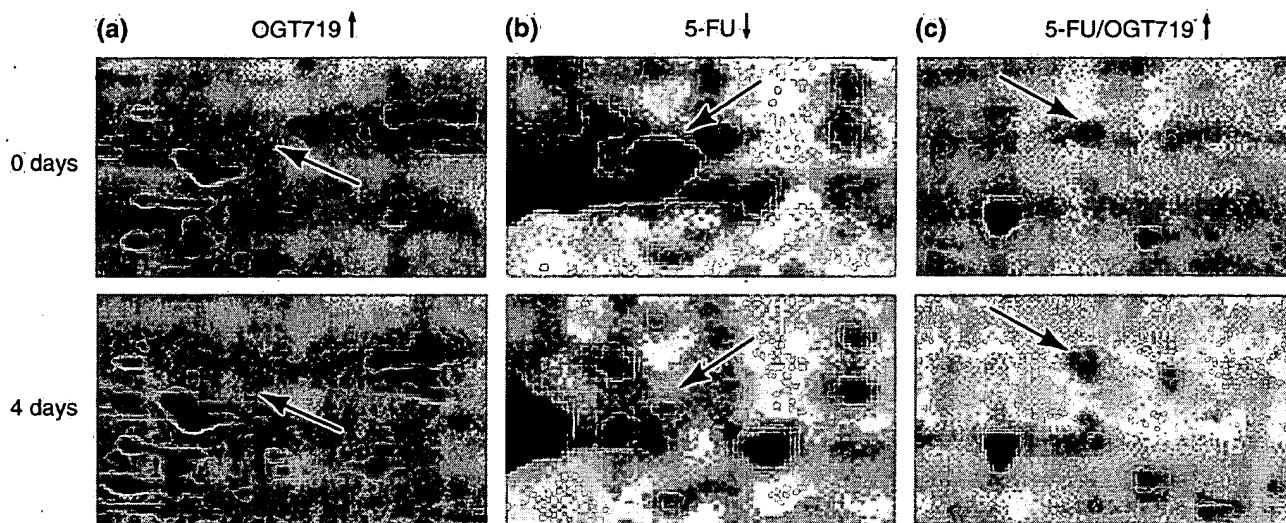
Figure 4. *Features that are specifically up- or downregulated in Huh7 cells by either 5-fluorouracil (5-FU) or OGT719: (a) elongation factor 1α2, (b) novel (three peptides by MS-MS) and (c) α-subunit of prolyl-4-hydroxylase. Arrows indicate up- or downregulated.*

of targeting, and being retained in, cells bearing the asialo-glycoprotein receptor (ASGP-r), including hepatocytes[29], hepatoma Huh7 cells[30] and some colorectal tumour cells[31]. The growth of the human hepatoma cell line Huh7 is inhibited by 5-FU or by OGT719. If the inhibition by OGT719 were the result of uptake and conversion to 5-FU as the active component, then it would be expected that Huh7 cells would show similar proteome profiles following exposure to either drug.

To examine these possibilities, we conducted an experiment taking samples of Huh7 cells that had been treated with IC_{50} doses of either OGT719 or 5-FU. Total cell lysates were prepared and taken through 2D electrophoresis, fluorescence staining, digital imaging and Proteograph analysis. To facilitate the interpretation of the data across all of the 2291 features seen on the proteomes, drug-induced protein changes of fivefold or greater, identified by the Proteograph, were analysed further. Interestingly, from this analysis 19 identical proteins were changed fivefold or more by both drugs, strongly suggesting similarities in the mode of action for these two compounds.

Thus, from very complex data involving >2000 protein features, using proteomics it is possible to analyse quantitatively and qualitatively each protein during its exposure to drugs. The biologist is now able to focus a series of further studies specifically on an enriched subset of proteins.

Figure 4 shows highlighted examples of the selected areas of the proteome where some of these identified proteins in the above study are altered in response to either or both drugs.

Several of the proteins identified above as being modulated similarly by 5-FU or OGT719 in Huh7 cells were subjected to tandem mass-spectrometric analysis for annotation. Some of these, such as the nuclear ribosomal RNA-binding protein[32], can be placed into pyrimidine pathways or related cell cycle/growth biochemical pathways in which 5-FU is known to act.

To attribute further significance to the proteome mode-of-action studies with OGT719, another cell line, the rat sarcoma HSN, was used. Growth of these cells is inhibited by 5-FU, but they are completely refractory to OGT719; notably they lack the ASGP-r, which might explain this finding (unpublished). For our proteome studies, HSN cells were treated with 5-FU or OGT719 over a time course of one, two and four days. At each time point, cells were harvested and processed to derive proteomes and Proteographs. As before, we purposely focused on those proteins that increased or decreased by fivefold or more. In this instance, there were no proteins co-modulated by the two drugs. This is perhaps to be expected, given that the HSN cells are killed by 5-FU and yet are refractory to OGT719.

Clear potential

The above is just an example of how proteomics can be used to address the mode of action of anticancer drugs. The potential of this approach is clear, and one can envisage situations where it will be profitable to compare the proteomes of cells in which the drug target has been eliminated by molecular knockout techniques, or with small-molecule inhibitors believed to act specifically on the same target. In addition to using proteomics to examine the action of drugs, it is also possible to use this approach to gauge the extent of nonspecific effects that might eventually lead to toxicity. For instance, in the example used above with HSN cells treated with OGT719, although cell growth was not affected, the levels of several specific proteins were changed. Further investigation of these proteins and the signalling pathways in which they are involved could be illuminating in predicting the likelihood or otherwise of long-term toxicity.

Us of proteomics in formal drug t xicology studies

A drug discovery programme at the stage where leads have been identified and mode-of-action studies are advanced, will proceed to investigate the pharmacokinetic and toxicology profile of those agents. These two parameters are of major importance in the drug discovery process, and many agents that have looked highly promising from in vitro studies have subsequently failed because of insurmountable pharmacokinetic and/or toxicity problems in vivo. Whereas the pharmacokinetic properties of a molecule can now be characterized quickly and accurately, toxicity studies are typically much longer and more demanding in their interpretation.

The ability to achieve fast and accurate predictions of toxicity within an in vivo setting would represent a big step forward in accelerating any drug discovery programme. Toxicity from a drug can be manifested in any organ. However, because the liver and kidney are the major sites in the body responsible for metabolism and elimination of most drugs, it is informative to examine these particular organs in detail to provide early indications about events that might result in toxicity.

The basis for most xenobiotic metabolizing activity is to increase the hydrophilicity of the compound and so facilitate its removal from the body. Most drugs are metabolized in the liver via the cytochrome P450 family of enzymes, which are known to comprise a total of ~200 different members[33,34], encompassing a wide array of overlapping specificities for different substrates. In addition to clearance, they also play a major role in metabolism that can lead to the production and removal of toxic species, and in some instances it is possible to correlate the ability or failure to remove such a toxin with a specific P450 or subgroup.

Unique P450 profiles

Each individual person will have a slightly different P450 profile, largely from polymorphisms and changes in expression levels, although other genetic and environmental factors aside from P450 also need to be taken into consideration. A significant amount of research is currently being directed towards this field – known as pharmacogenomics – with the aim of predicting how a patient will respond to a drug, as determined by their genetic make-up[35-37]. The marked variation of individuals in their ability to clear a compound can be one of the key factors in deciding the overall pharmacokinetic profile of a drug. Not only will this have a bearing on the likelihood of a patient responding to a treatment, but it will also be a factor in determining the possibility of their experiencing an adverse effect.

Many pharmaceutical companies are already employing genomic approaches, involving P450 measurements, as a key step in their assessment of the toxicological profile of a candidate drug and therefore of its suitability, or otherwise, to be considered for human clinical trials. There are limits to this approach, however. Whereas the P450 mRNA profiling can predict with some accuracy the likely metabolic fate of a drug, it will not provide information on whether the metabolites would subsequently lead to toxicity. Besides the patient-to-patient differences in steady-state levels of the P450s, there are also characteristic induction responses of these enzymes to some drugs. Moreover, as there can be some doubt over the correlation of mRNA levels and the corresponding protein levels, there is scope for misinterpretation of the results and hence real advantages to be gained from a proteome approach. In both instances, the ability to examine entire proteome profiles, including the P450 proteins, will be a significant advantage in understanding and predicting the metabolism and toxicological outcome of drugs.

In addition to direct organ and tissue studies, the serum, which collects the majority of toxicity markers released from susceptible organs and tissues throughout the entire body, can be utilized. Serum is rich in nuclease activity and, as pharmacogenomics is not suited to deal with these samples, valuable markers of toxicity could go undetected. However, by using proteomics for these types of analyses, serum markers (and clusters thereof) are now accessible for evaluation as indicators of toxicity.

Pharmacoproteomics

Proteomics can thus be used to add a new sphere of analysis to the study of toxicity at the protein level, and in the era of '-omics' there is a case to be made to adopt the term 'Pharmacoproteomics™'. Animals can be dosed with increasing levels of an experimental drug over time, and serum samples can be drawn for consecutive proteome analyses. Using this procedure, it should be possible to identify individual markers, or clusters thereof, that are dose related and correlate with the emergence and severity of toxicity. Markers might appear in the serum at a defined drug dose and time that are predictive of early toxicity within certain organs and if allowed to continue will have damaging consequences. These serum markers could subsequently be used to predict the response of each individual and allow tailoring of therapy whereby optimal efficacy is achieved without adverse side effects being apparent. This application can obviously extend to tracking toxicity of drugs in clinical trials where serum can be readily drawn and analysed. Surrogate markers for drug efficacy could also be detected by this procedure and could facilitate the challenge of identifying patient classes who will respond favourably to a drug and at what dosage.

Conclusions

By contrast to the agents administered to patients in clinical wards, the process of drug discovery is not a prescriptive series of steps. The risks are high and there are long timelines to be endured before it is known whether a candidate drug will succeed or fail. At each step of the drug discovery process there is often scope for flexibility in interpretation, which over many steps is cumulative. The pharmaceutical companies most likely to succeed in this environment are those that are able to make informed accurate decisions within an accelerated process.

The genomics revolution has impacted very positively upon these issues and now has a powerful new partner in proteomics. The ability to undertake global analysis of proteins from a very wide diversity of biological systems and to interrogate these in a high-throughput, systematic manner will add a significant new dimension to drug discovery. Each step of the process from target discovery to clinical trials is accessible to proteomics, often providing unique sets of data. Using the combination of genomics and proteomics, scientists can now see every dimension of their biological focus, from genes, mRNA, proteins and their subcellular localization. This will greatly assist our understanding of the fundamental mechanistic basis of human disease and allow new improved and speedier drug discovery strategies to be implemented.

REFERENCES

1   Crooke, S.T. (1998) *Nat. Biotechnol.* 16, 29–30
2   Dykes, C.W. (1996) *Br. J. Clin. Pharmacol.* 42, 683–695
3   Schena, M. *et al.* (1998) *Trends Biotechnol.* 16, 301–306
4   Ramsay, G. (1998) *Nat. Biotechnol.* 16, 40–44
5   Anderson, N.L. and Anderson, N.G. (1998) *Electrophoresis* 19, 1853–1861
6   James, P. (1997) *Biochem. Biophys. Res. Commun.* 231, 1–6
7   Wilkins, M.R. *et al.* (1996) *Biotechnol. Genet. Eng. Rev.* 13, 19–50
8   Parekh, R.B. and Rohlff, C. (1997) *Curr. Opin. Biotechnol.* 8, 718–723
9   Figeys, D. *et al.* (1998) *Electrophoresis* 19, 1811–1818
10 Wimmer, K. *et al.* (1996) *Electrophoresis* 17, 1741–1751
11 Giometti, C.S., Williams, K. and Tollaksen, S.L. (1997) *Electrophoresis* 18, 573–581
12 Williams, K. *et al.* (1998) *Electrophoresis* 19, 333–343
13 Rasmussen, R.K. *et al.* (1998) *Electrophoresis* 19, 818–825
14 Hirano, T. *et al.* (1995) *Br. J. Cancer* 72, 840–848
15 Ji, H. *et al.* (1997) *Electrophoresis* 18, 605–613
16 Ostergaard, M. *et al.* (1997) *Cancer Res.* 57, 4111–4117
17 Patel, V.B. *et al.* (1997) *Electrophoresis* 18, 2788–2794
18 Arnott, D. *et al.* (1998) *Anal. Biochem.* 258, 1–18
19 Anderson, L. and Seilhamer, J. (1997) *Electrophoresis* 18, 533–537
20 Rastan, S. and Beeley, L.J. (1997) *Curr. Opin. Genet. Dev.* 7, 777–783
21 Gravel, P. *et al.* (1995) *Electrophoresis* 16, 1152–1159
22 Qian, Y. *et al.* (1997) *Clin. Chem.* 43, 352–359
23 Sanchez, J.C. *et al.* (1997) *Electrophoresis* 18, 638–641
24 Watts, A.D. *et al.* (1997) *Electrophoresis* 18, 1086–1091
25 Asker, N. *et al.* (1995) *Biochem. J.* 308, 873–880
26 Ramsby, M.L., Makowski, G.S. and Khairallah, E.A. (1994) *Electrophoresis* 15, 265–277
27 Huber, L.A. (1995) *FEBS Lett.* 369, 122–125
28 Corthals, G.L. *et al.* (1997) *Electrophoresis* 18, 317–323
29 Hubbard, A.L., Wall, D.A. and Ma, A. (1983) *J. Cell Biol.* 96, 217–229
30 Zeng, F.Y., Oka, J.A. and Weigel, P.H. (1996) *Biochem. Biophys. Res. Commun.* 218, 325–330
31 Mu, J-Z. *et al.* (1994) *Biochim. Biophys. Acta* 1222, 483–491
32 Ghoshal, K. and Jacob, S.T. (1997) *Biochem. Pharmacol.* 53, 1569–1575
33 Guengerich, F.P. and Parikh, A. (1997) *Curr. Opin. Biotechnol.* 8, 623–628
34 Rendic, S. and Di Carlo, F.J. (1997) *Drug Metab. Rev.* 29, 413–580
35 Vermes, A., Guchelaar, H.J. and Koopmans, R.P. (1997) *Cancer Treat. Rev.* 23, 321–339
36 Housman, D. and Ledley, F.D. (1998) *Nat. Biotechnol.* 16, 492–493
37 Persidis, A. (1998) *Nat. Biotechnol.* 16, 209–210

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF TOD BEDILION, Ph.D.
UNDER 37 C.F.R. § 1.132

I, TOD BEDILION, Ph.D., declare and state as follows:

1. In April, 1996, I became the first employee of Synteni, Inc., where I served as Research Director until its acquisition by Incyte Corporation in early 1998. After Synteni's acquisition, I continued in the position of Director of Corporate Development at Incyte until May 11, 2001. I am currently the Director of Business Development at Genomic Health, Inc., Redwood City, California and an occasional Consultant to Incyte.

2. Synteni was founded to commercialize expression microarrays, microarrays in which expressed nucleic acids -- full-length cDNAs, fragments of full-length cDNAs, expressed sequence tags (ESTs) -- are arrayed on a common support to permit highly parallel detection and measurement of the expression of their cognate genes in a biological sample.

3. During my employ at Synteni, virtually all (if not all) of my work efforts were directed to the further technical development and the commercial exploitation of that microarray technology; given the small size of our shop, most of us had both technical and commercial responsibilities. The customer accounts for which I was personally responsible included large pharmaceutical companies, such as SmithKline

Beecham, large biotechnology companies, such as Genentech, and small research institutes, such as DNAX Inc.

4.   From my very first interaction with our customers, consistently through to Synteni's acquisition by Incyte, I heard uniform, consistent, and emphatic requests that more genes be added to the arrays. This was true with respect to both our original microarrays, based on customer-provided genes and libraries, and our later, "generic", gene expression microarrays, based upon the unigene clone collection (our so-called "UniGem" arrays). From day 1, the pressure on us was to print ever more spots on the array. It was **never** a question: our customers wanted ever more genes on the array, each new gene-specific probe providing incrementally more value to the customer.[1]

5.   As a commercial enterprise, providing value to our customers was our major concern. Thus, to increase the value of our products and services in the marketplace -- to increase our ability to sell our microarrays and microarray services, their "salability" -- our efforts from the very beginning were devoted to increasing the number of specific genes whose expression could be detected with our microarrays.

6.   Indeed, one of our major competitive advantages in the marketplace -- not just as regards other commercial suppliers, but also with respect to the innumerable laboratories and companies that were attempting to spot arrays in their own "home-brew" facilities -- was the number of

---

[1]   I should note the customers were not asking for addition of probes specific to only those genes for which the biological function of the encoded gene product was known, but were asking for probes specific to any and all expressed genes.
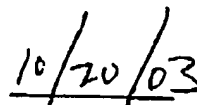
distinct gene-specific probes that we provided on our expression microarrays. Our first 10,000 element UniGem array put the holy grail of gene expression analysis -- the human whole genome array -- within sight for the very first time (with respect to timing of the UniGEM program we began project planning and technology development in mid 1996 and delivered our first 10,000 element standard content human arrays in the first months of 1997 as I recall).

7.   By the end of 1997, our efforts to provide the most comprehensive, and thus most valuable, human gene expression microarrays had been sufficiently successful that Incyte agreed to acquire Synteni for a reported $80 million.

8.   I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.

_____
Tod Bedilion, Ph.D.

10/20/03
Date

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

DECLARATION OF VISHWANATH R. IYER, Ph.D.
UNDER 37 C.F.R. § 1.132

I, VISHWANATH R. IYER, Ph.D., declare and state as follows:

1.   I am an Assistant Professor in the Section of Molecular Genetics and Microbiology, Institute of Cellular and Molecular Biology, University of Texas at Austin, where my laboratory currently studies global transcriptional control in yeast, gene expression programs during human cell proliferation, and genome-wide transcription factor targets in yeast and human. Immediately prior to this position, I spent four years as a postdoctoral fellow in the laboratory of Patrick O. Brown at Stanford University studying the transcriptional programs of yeast and of human cells. My *curriculum vitae* is attached hereto as Exhibit A.

2.   Beginning in Dr. Brown's laboratory, where I helped to develop the first whole genome arrays for yeast and early versions of highly representative cDNA arrays for human cells, and continuing to the present day, I have used microarray-based gene expression analysis as a principal approach in much of my research.

3.   Representative publications describing this work include:

DeRisi J. *et al.*, "Exploring the metabolic and genetic control of gene expression on a genomic scale," *Science* 278:680-686 (1997);[1]

Marton *et al.*, "Drug target validation and identification of secondary drug target effects using DNA microarrays," Nature Med. 4:1293-1301 (1998);[2]

Iyer *et al.*, "The transcriptional program in the response of human fibroblasts to serum," Science 283:83-87 (1999);[3] and

Ross *et al.*, "Systematic variation in gene expression patterns in human cancer cell lines," *Nature Genetics* 24: 227-235 (2000).[4]

Two of the papers describe our use of microarray-based expression profiling to explore the metabolic reprogramming that occurs during major environmental changes, both in yeast (DeRisi *et al.*, during the shift from fermentation to respiration) and in human cells (Iyer *et al.*, human fibroblasts exposed to serum). One reference describes our use of expression profile analysis in drug target validation and identification of secondary drug effects (Marton *et al.*). And one describes our use of expression profiling as a molecular phenotyping tool to discriminate among human cancer cells (Ross *et al.*).

4.   Whether used to elucidate basic physiological responses, to study primary and secondary drug effects, or to discriminate and classify human cancers, expression profiling

---

[1]   Attached hereto as Exhibit B.
[2]   Attached hereto as Exhibit C.
[3]   Attached hereto as Exhibit D.
[4]   Attached hereto as Exhibit E.

as we have practiced it relies for its power on comparison of **patterns** of expression.

5. For example, we have demonstrated that we can use the presence or absence of a characteristic drug "signature" pattern of altered gene expression in drug-treated cells to explore the mechanism of drug action, and to identify secondary effects that can signal potentially deleterious drug side effects. As another example, we have demonstrated that gene expression patterns can be used to classify human tumor cell lines. While it is of course advantageous to know the biological function of the encoded gene products in order to reach a better understanding of the cellular mechanisms underlying these results, these pattern-based analyses do not require knowledge of the biological function of the encoded proteins.

6. The resolution of the patterns used in such comparisons is determined by the number of genes detected: the greater the number of genes detected, the higher the resolution of the pattern. It goes without saying that higher resolution patterns are generally more useful in such comparisons than lower resolution patterns. With such higher resolutions comes a correspondingly higher degree of statistical confidence for distinguishing different patterns, as well as identifying similar ones.

7. Each gene included as a probe on a microarray provides a signal that is specific to the cognate transcript, at least to a first approximation.[5] Each new gene-specific

---

[5] In a more nuanced view, it is certainly possible for a probe to signal the presence of a variety of splice variants of a single gene,

(Continued...)

-3-

probe added to a microarray thus increases the number of genes detectable by the device, increasing the resolving power of the device. As I note above, higher resolution patterns are generally more useful in comparisons than lower resolution patterns. Accordingly, each new gene probe added to a microarray increases the usefulness of the device in gene expression profiling analyses. This proposition is so well-established as to be virtually an axiom in the art, and has been as long as I have been working in the field, and certainly since the time I embarked on the production of whole genome arrays in early 1996. Simply put, arrays with fewer gene-specific probes are inferior to arrays with more gene-specific probes.

8. For example, our ability to subdivide cancers into discriminable classes by expression profiling is limited by the resolution of the patterns produced. With more genes contributing to the expression patterns, we can potentially draw finer distinctions among the patterns, thus subdividing otherwise indistinguishable cancers into a greater number of classes; the greater the number of classes, the greater the likelihood that the cancers classified together will respond similarly to therapeutic intervention, permitting better individualization of therapy and, we hope, better treatment outcomes.

9. If a gene does not change expression in an experiment, or if a gene is not expressed and produces no

_____

(_Continued)
without discriminating among them, and for a probe to signal the presence of a variety of allelic variants of a single gene, again without discriminating among them.

signal in an experiment, that is not to say that the probe lacks usefulness on the array; it only means that an insufficient number of conditions have been sampled to identify expression changes. In fact, an experiment showing that a gene is not expressed or that its expression level does not change can be equally informative. To provide maximum versatility as a research tool, the microarray should include -- and as a biologist I would want my microarray to include -- each newly identified gene as a probe.

10. I declare further that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true, and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and may jeopardize the validity of any patent application in which this declaration is filed or any patent that issues thereon.

_____
VISHWANATH R. IYER, Ph.D.

October 20, 2003
Date

Vishwanath R. Iyer

Assistant Professor
Section of Molecular Genetics and Microbiology
Institute of Cellular and Molecular Biology
MBB 3.212A, University of Texas at Austin
Austin, TX 78712-0159
Phone: 512-232-7833
Fax: 512-232-3432
Email: vishy@mail.utexas.edu

Education/Training

Bombay University, Mumbai, India	B.Sc. (1987), Chemistry & Biochemistry
M. S. University of Baroda, Baroda, India	M.Sc. (1989), Biotechnology
Harvard University, Cambridge MA	Ph.D. (1996), Genetics
Stanford University, Stanford CA	Post-doctoral (1996-2000), Genomics

Research Experience

9/00-5/03   Assistant professor, Section of Molecular Genetics and
                Microbiology, University of Texas, Austin TX
- Global transcriptional control in yeast
- Gene expression programs during human cell proliferation
- Genome-wide transcription factor targets in yeast and human
- Collaborative microarray facility

5/96-8/00   Post-doctoral fellow   Stanford University, Stanford CA
                (Advisor: Dr. Patrick O. Brown)
- Yeast whole-genome ORF and intergenic microarrays
- Human cDNA microarrays for expression profiling

9/89-4/96   Graduate student   Harvard University, Cambridge MA
                (Advisor: Dr. Kevin Struhl)
- Yeast transcriptional regulation

Honours and Awards

Government of India Biotechnology Fellowship (1987-1989)
University Grants Commission Junior Research Fellowship (1989)
Stanford University/NHGRI Genome Training Grant (1996)

Invited Conference talks (selected)

Invited Lecturer, NEC-Princeton Lectures in Biophysics
    Princeton, NJ (June 1998)
Plenary Session Speaker, HGM '99 (HUGO Human Genome Meeting)
    Brisbane, Australia (April 1999)
Invited Speaker, Gordon Research Conference "Human Molecular Genetics"
    Newport, RI (August 2001)

Invited Speaker, Nature Genetics "Oncogenomics 2002" Conference
Dublin, Ireland (May 2002)
Invited Speaker, "Pathology Bioinformatics" Symposium, University of Michigan,
Ann Arbor, MI (November 2002)
Invited Speaker, "Systems Biology: Genomic Approaches to Transcriptional
Regulation" Cold Spring Harbor Laboratory Meeting (March 2003)
Symposium co-Chair and Speaker "Functional Genomics" American Society for
Biochemistry and Molecular Biology Meeting, San Diego, CA (April 2003)
Invited Speaker in Functional Genomics (Gene Networks) Symposium, International
Congress of Genetics, Melbourne Australia July 6-11 2003
Invited Speaker "BioArrays Europe 2003"
Cambridge, UK (Sep/Oct 2003)

Departmental Seminars
Texas A&M University Genetics and Biochemistry & Biophysics Departments,
October 24 2002
New York University School of Medicine, Department of Biochemistry,
November 20 2002
UT Southwestern Medical Center, Human Genetics Seminar Series,
May 5 2002
UCLA School of Medicine, Department of Human Genetics
June 2 2003
National Human Genome Research Institute
June 12 2003
Sanger Institute of the Wellcome Trust, Hinxton, UK
Sep 2003

Other Professional Activities
Reviewer for *Genome Biology, Genome Research, Nature Genetics, Science* (1998-2003)
Instructor, Cold Spring Harbor Summer Course "Making and using DNA Microarrays"
(2000 - 2003)
Member, NIDDK Special Emphasis Review Panel ZDK1 (2001-2002)

Publications
1. Iyer V. & Struhl, K. (1995) Poly(dA:dT), a ubiquitous promoter element that
stimulates transcription via its intrinsic DNA structure, *EMBO J.* 14: 2570-2579.

2. Iyer V. & Struhl, K. (1995) Mechanism of differential utilization of the his3 TR and TC
TATA elements, *Mol. Cell. Biol.* 15: 7059-7066.

3. Iyer V. & Struhl K. (1996) Absolute mRNA levels and transcription initiation rates in
Saccharomyces cerevisiae. Proc. Natl. Acad. Sci . (USA) 93:5208-5212.

4. DeRisi J. L., Iyer V. R. & Brown P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278:680-686

5. Marton M. J., DeRisi J. L., Bennett H. A., Iyer V. R., Meyer M. R., Roberts C. J., Stoughton R., Burchard J., Slade D., Dai H., Bassett D. E. Jr., Hartwell L. H., Brown P. O. & Friend S. H. (1998) Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nature Med.* 4:1293-1301

6. Lutfiyya L. L., Iyer V. R., DeRisi J., DeVit M. J., Brown P. O. & Johnston M. (1998) Characterization of three related glucose repressors and genes they regulate in *Saccharomyces cerevisiae*. *Genetics* 150:1377-1391

7. Spellman P. T., Sherlock G., Zhang M. Q., Iyer V. R., Anders K., Eisen M. B., Brown P. O., Botstein D. & Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell* 9:3273-3297

8. Iyer V. R., Eisen M. B., Ross D. T., Schuler G., Moore T., Lee J. C.,F., Trent J. M., Staudt L. M., Hudson Jr. J., Boguski M. S., Lashkari D., Shalon D., Botstein D. & Brown P. O. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science* 283:83-87

9. DeRisi J. L. & Iyer V. R. (1999) Genomics and array technology. *Curr. Opin. Oncol.* 11:76-79

10. Ross D. T., Scherf U., Eisen M. B., Perou C. M., Spellman P., Iyer V. R., Rees C., Jeffrey S. S., Van de Rijn M., Waltham M., Pergamenschikov A., Lee J. C. F., Lashkari D., Shalon D., Myers T. G., Weinstein J. N., Botstein D., & Brown P. O. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics* 24: 227-235

11. Sudarsanam P., Iyer V. R., Brown P. O. & Winston F. (2000) Whole-genome expression analysis of *snf/swi* mutants of *S. cerevisiae*. *Proc. Natl. Acad. Sci .(USA)* 97: 3364-3369

12. Tran H. G., Steger D. J., Iyer V. R., & Johnson A. D. (2000) The chromo domain protein Chd1p from budding yeast is an ATP-dependent chromatin-modifying factor *EMBO J* 19: 2323-2331

13. Gross C., Kelleher M., Iyer V. R., Brown P. O., & Winge D. R.. (2000) Identification of the copper regulon in *Saccharomyces cerevisiae* by DNA microarrays. *J. Biol. Chem.* 275: 32310-32316

14. Reid J. L., Iyer V. R., Brown P. O. & Struhl K. (2000) Coordinate regulation of yeast ribosomal protein genes is associated with targeted recruitment of Esa1 histone acetylase. *Mol. Cell* 6: 1297–1307

15. Iyer V. R., Horak C., Scafe C. S., Botstein D., Snyder M. & Brown P. O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF *Nature* 409: 533-538

16. Miki R., Kadota K., Bono H., Mizuno Y., Tomaru Y., Carninci P., Itoh M., Shibata K., Kawai J., Konno H., Watanabe S., Sato K., Tokusumi Y., Kikuchi N., Ishii Y., Hamaguchi Y., Nishizuka I., Goto H., Nitanda H., Satomi S., Yoshiki A., Kusakabe M., DeRisi J.L., Eisen M.B., Iyer V.R., Brown P.O., Muramatsu M., Shimada H., Okazaki Y. & Hayashizaki Y. (2001) Delineating developmental and metabolic pathways in vivo by expression profiling using the RIKEN set of 18,816 full-length enriched mouse cDNA arrays *Proc. Natl. Acad. Sci. (USA)* 98: 2199-2204

17. Pollack J. R. & Iyer V.R. (2002) Characterizing the physical genome. *Nature Genetics* 32 suppl: 515-521

18. Iyer V. R. Microarray-based detection of DNA protein interactions: Chromatin Immunoprecipitation on Microarrays, in *DNA Microarrays: A Molecular Cloning Manual* (eds. Bowtell, D. & Sambrook, J.) 453-463 (Cold Spring Harbor Laboratory Press, 2003).
(not peer reviewed)

19. Killion, P., Sherlock G. and Iyer V. R. (2003) The Longhorn Array Database, an open-source implementation of the Stanford Microarray Database *BMC Bioinformatics* 4: 32

20. Hahn J. S., Hu Z., Thiele D. J. & Iyer V. R. Genome-Wide Analysis of the Biology of Stress Responses Through Heat Shock Transcription Factor (submitted to *PNAS*)

21. Kim J. & Iyer V.R. The global role of TBP recruitment to promoters in mediating gene expression profiles (manuscript in preparation)

Current/Pending Research Support

003658-0223-2001 Iyer (PI) 16% effort
01/01/02 - 08/31/04
Texas Higher Education Coordinating Board (ARP)
"Microarray based global mapping of DNA-protein interactions at promoters in human cells"
This is a pilot project to map the in vivo interactions of transcription factors with human promoters
Role: PI

Information Technology Research 0325116 R. Mooney (PI) 9% effort
09/01/03 - 08/31/07
NSF
"Feedback from Multi-Source Data Mining to Experimentation for Gene Network Discovery"
Role: Co-investigator

1 R01 CA95548-01A2 (pending) Iyer (PI) 25% effort
12/1/03 – 11/30/08
NIH
"Analysis of genome-wide transcriptional control in yeast"
This is a project to identify stress responsive transcription factor targets in yeast through the use of DNA microarrays
Role: PI

Breast Cancer Idea Award (pending) Iyer (PI) 10% effort
1/1/04 – 12/31/06
US Army Medical Research and Materiel Command
"Genome-wide chromosomal targets of oncogenic transcription factors"
This is a project aimed at identifying direct chromosomal targets of c-myc and ER in human cells through the use of a novel sequence tag analysis method.
Role: PI
003658-0531-2003 (pending) Marcotte (PI) 8% effort
01/01/04 - 12/31/05
Texas Higher Education Coordinating Board (ATP)
"Cell arrays: A novel high–throughput platform for measuring gene function on a genomic scale"
This proposal is aimed at developing a novel microarray based platform for automated, high–throughput microscopic imaging of cells, allowing rapid and systematic evaluation of gene function.

Fischer-Vize, Science 270, 1828 (1995).
35. T. C. James and S. C. Elgin, Mol. Cell Biol. 6, 3862 (1986); R. Paro and D. S. Hogness, Proc. Natl. Acad. Sci. U.S.A. 88, 263 (1991); B. Tschiersch et al., EMBO J. 13, 3822 (1994); M. T. Madreddi et al., Cell 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, Proc. Natl. Acad. Sci. U.S.A. 93, 7137 (1996).
36. P. M. Palosaari et al., J. Biol. Chem. 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E.

Grund, R. Eichenlaub, Appl. Environ. Microbiol. 58, 4068 (1992); V. Sharma, K. Suvama, R. Meganathan, M. E. Hudspeth, J. Bacteriol. 174, 5057 (1992); M. Kanazawa et al., Enzyme Protein 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, J. Bacteriol. 178, 3015 (1996).
37. M. Ho et al., Cell 77, 869 (1994).
38. W. Hendriks et al., J. Cell Biochem. 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of Saccharomyces cerevisiae were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns f many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by d letion of the transcriptional co-repressor TUP1 or overexpression of the transcriptional activator YAP1. These results demonstrate the feasibility and utility of this approach t genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (1, 2), provide a practical and economical tool for studying gene expression on a very large scale (3–6).

Saccharomyces cerevisiae is an especially

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (7). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (8). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (9). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (10). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (11) and then hybridized to the microarrays (12). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (13).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (14). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

t any gene whose function is known (15). The responses of these previously unchar-acterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression f genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl–coenzyme A(CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes PCK1, encoding phosphoenolpyruvate carboxykinase, and FBP1, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, ACR1 and IDP2, revealed that ACR1, a gene essential for ACS1 activity, also possessed a consensus CSRE motif, but interestingly, IDP2 did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception
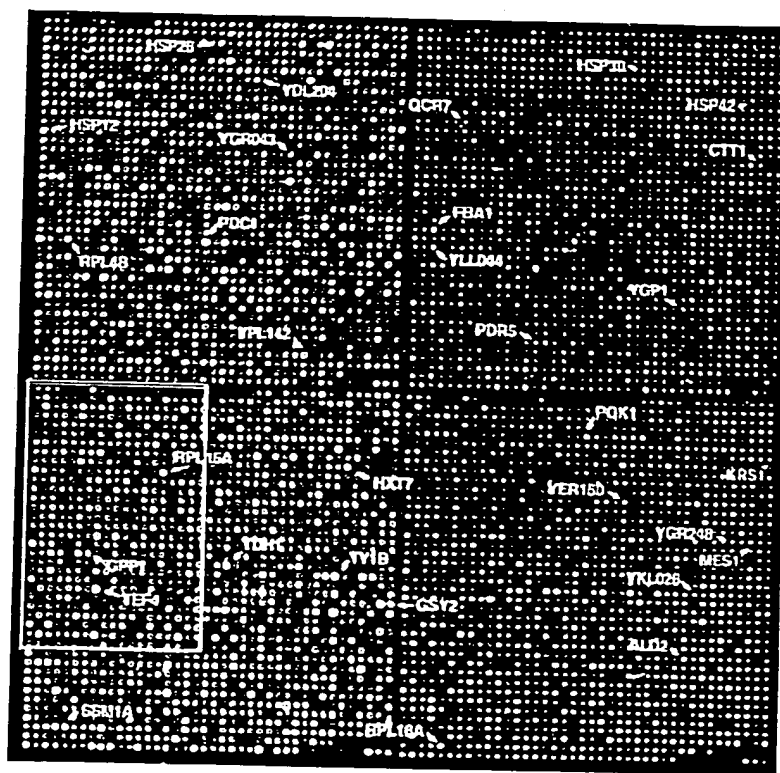

Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of <5 × 10^6 cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of ~2 × 10^8 cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been sh wn to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream f HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS$_{rpg}$) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of
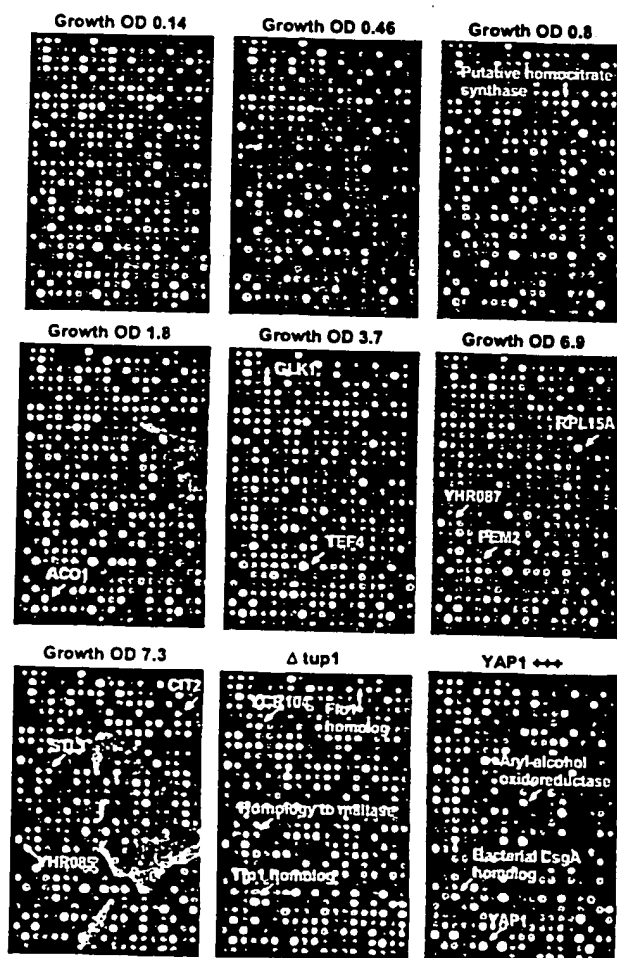
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

si n ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, ther were discrepancies between our results and previous results, pointing t technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the tup1Δ mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (39). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (40).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1Δ*) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (11). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1Δ* strain, and thus presumably repressed by Tup1 (41). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1Δ* mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (13)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1Δ* strain, providing a positive control in the experiment (42).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (15) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (43), and similar, serine-poor proteins, the seripauperins (44). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1Δ*
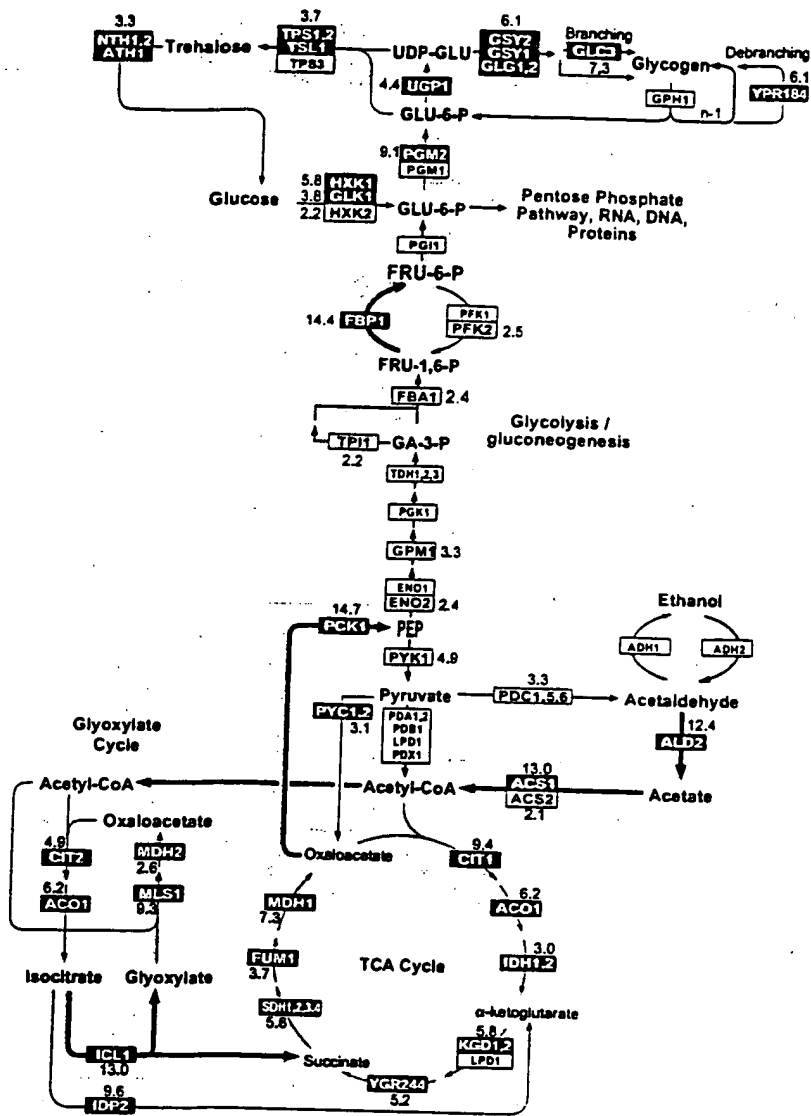

Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when TUP1 was deleted. In contrast, none f 83 genes that could be classified as putative regulators f the cell divisi n cycle w re induced more than twofold by deletion of TUP1. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monit r expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets f a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a MAT α strain in which MFA1 and MFA2, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic tup1Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a MATA strain (in which expression of MFA1 and MFA2 is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. YAP1 encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of YAP1 in yeast confers increased resistance to hydrogen peroxide, o-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing YAP1 under the control of the strong GAL1-10 promoter, both grown in galactose (that is, a condition that induces YAP1 overexpression). Complementary DNA from the control and YAP1 overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing YAP1.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in S. cerevisiae, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain verexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxy-late-cycle enzymes, 24.
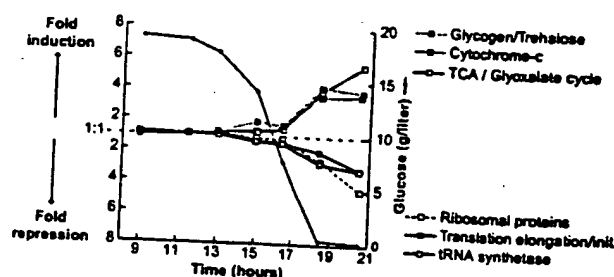

Table 1. Genes induced by YAP1 overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon YAP1 overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W		YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to A. thaliana zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with ther factors, or less likely, may represent an indirect effect f Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists f components that were chosen for their modest cost and simplicity. It was feasible f r a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Science 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, Genome Res. 6, 639 (1996).
3. D. Lashkari, Proc. Natl. Acad. Sci. U.S.A., in press.
4. J. DeRisi et al., Nature Genet. 14, 457 (1996).
5. D. J. Lockhart et al., Nature Biotechnol. 14, 1675 (1996).
6. M. Chee et al., Science 274, 610 (1996).
7. M. Johnston and M. Carlson, in The Molecular Biology of the Yeast Saccharomyces: Gene Expression, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-μl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3x standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-

Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of IDP2, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) SAM1, GPP1, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251). Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.

11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)+] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).

12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.

13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html

14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.

15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database (genome-www.stanford.edu), Yeast Protein Database (quest7.proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/mips/yeast/index.htmlx).

16. A. Scholer and H. J. Schuller, Mol. Cell. Biol. 14, 3613 (1994).

17. S. Kratzer and H. J. Schuller, Gene 161, 75 (1995).

18. R. J. Haselbeck and H. L. McAlister, J. Biol. Chem. 268, 12116 (1993).

19. M. Fernandez, E. Fernandez, R. Rodicio, Mol. Gen. Genet. 242, 727 (1994).

20. A. Hartig et al., Nucleic Acids Res. 20, 5677 (1992).

21. P. M. Martinez et al., EMBO J. 15, 2227 (1996).

22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, Mol. Cell. Biol. 15, 6232 (1995).

23. H. Ruis and C. Schuller, Bioessays 17, 959 (1995).

24. L. Parrou, M. A. Teste, J. Francois, Microbiology 143, 1891 (1997).

25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.

26. S. L. Forsburg and L. Guarente, Genes Dev. 3, 1166 (1989).

27. J. T. Olesen and L. Guarente, ibid. 4, 1714 (1990).

28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, Mol. Microbiol. 13, 119 (1994).

29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B—C, G, or T; N—G, A, T, or C; R—A or G; and Y—C or T.

30. C. Fondrat and A. Kalogeropoulos, Comput. Appl. Biosci. 12, 363 (1996).

31. D. Shore, Trends Genet. 10, 408 (1994).

32. R. J. Planta and H. A. Raue, ibid. 4, 64 (1988).

33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.

34. S. F. Neuman, S. Bhattacharya, J. R. Broach, Mol. Cell. Biol. 15, 3187 (1995).

35. P. Lesage, X. Yang, M. Carlson, ibid. 16, 1921 (1996).

36. For example, we observed large inductions of the genes coding for PCK1, FBP1 [Z. Yin et al., Mol. Microbiol. 20, 751 (1996)], the central glyoxylate cycle gene ICL1 [A. Scholer and H. J. Schuller, Curr. Genet. 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, ACS1 [M. A. van den Berg et al., J. Biol. Chem. 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes PYK1 and PFK2 [P. A. Moore et al., Mol. Cell. Biol. 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T CTT1 [P. H. Bissinger et al., ibid. 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as HSP12, HSP26, and HSP42 [I. Farkas et al., J. Biol. Chem. 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, Mol. Gen. Genet. 223, 97 (1990); D. Wotton et al., J. Biol. Chem. 271, 2717 (1996)].

37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, FBP1 and PCK1) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).

38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, ADH1 and ADH2, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as HXK1/HXK2 (77% identical) [P. Herrero et al., Yeast 11, 137 (1995)], MLS1/DAL7 (73% identical) (20), and PGM1/PGM2 (72% identical) [D. Oh, J. E. Hopper, Mol. Cell. Biol. 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.

39. F. E. Williams, U. Varanasi, R. J. Trumbly, Mol. Cell. Biol. 11, 3307 (1991).

40. D. Tzamarias and K. Struhl, Nature 369, 758 (1994).

41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concor-

dance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.

42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of TUP1 and an Eco R I site 124 base pairs before the stop codon of the TUP1 gene.

43. L. R. Kowalski, K. Kondo, M. Inouye, Mol. Microbiol. 15, 341 (1995).

44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, Gene 148, 149 (1994).

45. D. Hirata, K. Yano, T. Miyakawa, Mol. Gen. Genet. 242, 250 (1994).

46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, Appl. Environ. Microbiol. 60, 1783 (1994).

47. A. Muheim et al., Eur. J. Biochem. 195, 369 (1991).

48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, J. Biol. Chem. 269, 32592 (1994).

49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm. stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.

50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold)

51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

Drug target validation and identification of secondary drug target effects using DNA microarrays

MATTHEW J. MARTON[1], JOSEPH L. DERISI[2], HOLLY A. BENNETT[1], VISHWANATH R. IYER[2], MICHAEL R. MEYER[1], CHRISTOPHER J. ROBERTS[1], ROLAND STOUGHTON[1], JULJA BURCHARD[1], DAVID SLADE[1], HONGYUE DAI[1], DOUGLAS E. BASSETT, JR[1]., LELAND H. HARTWELL[3], PATRICK O. BROWN[2] & STEPHEN H. FRIEND[1]

[1]Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034, USA
[2]Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute Stanford, California 94305-5428, USA
[3]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue N., Seattle, Washington 98109, USA
Correspondence should be addressed to S.H.F; email: sfriend@rosetta.org

We describe here a method for drug target validation and identification of secondary drug target effects based on genome-wide gene expression patterns. The method is demonstrated by several experiments, including treatment of yeast mutant strains defective in calcineurin, immunophilins or other genes with the immunosuppressants cyclosporin A or FK506. Presence or absence of the characteristic drug 'signature' pattern of altered gene expression in drug-treated cells with a mutation in the gene encoding a putative target established whether that target was required to generate the drug signature. Drug dependent effects were seen in 'targetless' cells, showing that FK506 affects additional pathways independent of calcineurin and the immunophilins. The described method permits the direct confirmation of drug targets and recognition of drug-dependent changes in gene expression that are modulated through pathways distinct from the drug's intended target. Such a method may prove useful in improving the efficiency of drug development programs.
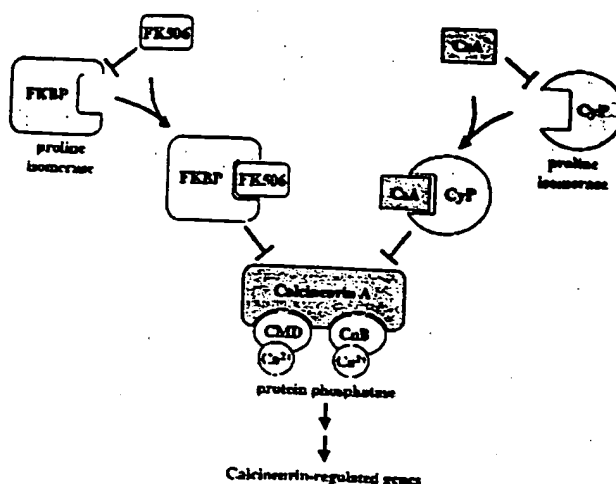
Good drugs are potent and specific; that is, they must have strong effects on a specific biological pathway and minimal effects on all other pathways. Confirmation that a compound inhibits the intended target (drug target validation) and the identification of undesirable secondary effects are among the main challenges in developing new drugs. Comprehensive methods that enable researchers to determine which genes or activities are affected by a given drug might improve the efficiency of the drug discovery process by quickly identifying potential protein targets, or by accelerating the identification of compounds likely to be toxic. DNA microarray technology, which permits simultaneous measurement of the expression levels of thousands of genes, provides a comprehensive framework to determine how a compound affects cellular metabolism and regulation on a genomic scale[1-11]. DNA microarrays that contain essentially every open reading frame (ORF) in the *Saccharomyces cerevisiae* genome have already been used successfully to explore the changes in gene expression that accompany large changes in cellular metabolism or cell cycle progression[7-10].

In the modern drug discovery paradigm, which typically begins with the selection of a single molecular target, the ideal inhibitory drug is one that inhibits a single gene product so completely and so specifically that it is as if the gene product were absent. Treating cells with such a drug should induce changes in gene expression very similar to those resulting from deleting the gene encoding the drug's target. Here we have compared the genome-wide effects on gene expression that result from deletions of various genes in the budding yeast *S. cerevisiae* to the effects on gene expression that result from treatment

with known inhibitors of those gene products. Using the calcineurin signaling pathway as a model system, we tested an approach that permits identification of genes that encode proteins specifically involved in pathways affected by a drug. The FK506 characteristic pattern, or 'signature', of altered gene expression was not observed in mutant cells lacking proteins inhibited by FK506 (for example, a calcineurin or FK506-binding-protein mutant strain), but was observed in mutants deleted for genes in pathways unrelated to FK506 action (for example, a cyclophilin mutant strain). Conversely, the cyclosporin A (CsA) signature was not observed in CsA-treated calcineurin or cyclophilin mutant strains, but was seen in an FK506-binding-protein mutant strain treated with CsA. The method also demonstrates that FK506, a clinically used immunosuppressant, has 'off-target' effects that are independent of its binding to immunophilins. Thus, the approach we describe may provide a way to identify the pathways altered by a drug and to detect drug effects mediated through unintended targets.

Null mutants phenocopy drug-treated cells on a genomic scale

To test whether a null mutation in a drug target serves as a model of an ideal inhibitory drug, we examined the effects on gene expression associated with pharmacological or genetic inhibition of calcineurin function. Calcineurin is a highly conserved calcium- and calmodulin-activated serine/threonine protein phosphatase implicated in diverse processes dependent on calcium signaling[12-13]. In budding yeast, calcineurin is required for intracellular ion homeostasis[14], for adaptation to prolonged mating pheromone treatment[15] and in the regulation of

Fig. 1 Model of antagonism of the calcineurin signaling pathway mediated by FK506 and cyclosporin A (CsA). Calcineurin activity is composed of a catalytic subunit (calcineurin A, encoded in yeast by the *CNA1* and *CNA2* genes), and calcium-binding regulatory subunits calmodulin (CMD) and calcineurin B (CnB). After entering cells, FK506 and CsA specifically bind and inhibit the peptidyl-proline isomerase activity of their respective immunophilins. FK506 binding proteins (FKBP) and cyclophilins (CyP). The most abundant immunophilins in yeast (Fpr1 and Cph1) are thought to mediate calcineurin inhibition. Drug-immunophilin complexes bind and inhibit the calcium- and calmodulin-stimulated phosphatase calcineurin. Among the substrates of calcineurin are transcriptional activators that act to modulate gene expression.

the onset of mitosis[16]. In mammals, calcineurin has been implicated in T-cell activation[12], in apoptosis[17], in cardiac hypertrophy[18] and in the transition from short-term to long-term memory[19]. In both organisms, calcineurin activity is inhibited by FK506 and CsA, immunosuppressant drugs whose effects on calcineurin are mediated through families of intracellular receptor proteins called immunophilins[12,20] (Fig. 1). To assess the effects of pharmacologic inhibition of calcineurin, wild-type *S. cerevisiae* was grown to early logarithmic phase in the presence or absence of FK506 or CsA. Isogenic cells, from which the genes encoding the catalytic subunits of calcineurin (*CNA1* and *CNA2*) had been deleted[21] (referred to as the *cna* or calcineurin mutant), were grown in parallel, in the absence of the drug. Fluorescently-labeled cDNA was prepared by reverse transcription of polyA+ RNA in the presence of Cy3- or Cy5-deoxynucleotide triphosphates and then hybridized to a microarray containing more than 6,000 DNA probes representing 97% of the known or predicted ORFs in the yeast genome. Simultaneous hybridization of Cy5-labeled cDNA from mock-treated cells and Cy3-labeled cDNA from cells treated with 1 µg/ml FK506 allowed the effect of drug treatment on mRNA levels of each ORF to be determined (Fig. 2a and b and data not shown). Similarly, effects of the calcineurin mutations on the mRNA levels of each gene were assessed by simultaneous hybridization of Cy5-labeled cDNA from wild-type cells and Cy3-labeled cDNA from the calcineurin mutant strain(Fig. 2c). For each comparison of this kind, reported expression ratios are the average of at least two hybridizations in which the Cy3 and Cy5 fluors were reversed to remove biases that may be introduced by gene-specific differences in incorporation of the two fluors (data not shown).
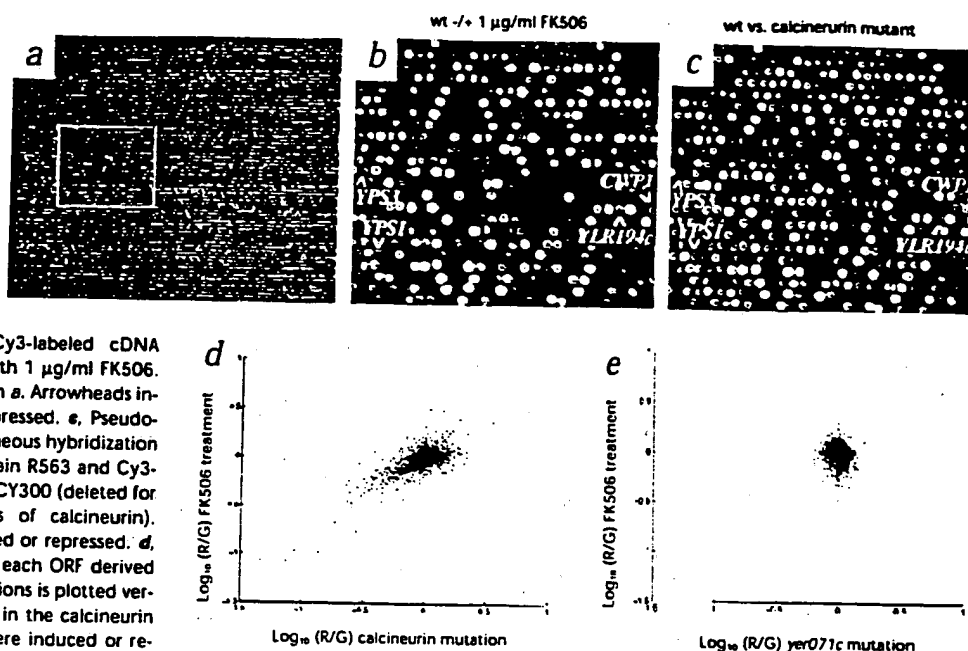
Treatment with FK506 in these growth conditions resulted in a signature pattern of altered gene expression in which mRNA levels of 36 ORFs changed by more than twofold (http://www.rosetta.org). A very similar pattern of altered gene expression was observed when the calcineurin mutant strain was compared to wild-type cells. Comparison of the changes in mRNA expression of each gene resulting from treatment of wild-type cells with FK506 with mRNA expression changes resulting from deletion of the calcineurin genes showed the considerable similarity of the global transcript alterations in response to the two perturbations (Fig. 2b–d). Quantification of this similarity using the correlation coefficient (ρ) showed large correlations between the FK506 treatment signature and the calcineurin deletion signature (ρ = 0.75 ± 0.03), as well as the CsA treatment signature (ρ = 0.94±0.02), but not with a randomly selected deletion mutant strain (deleted for the *YER071C* gene; ρ = -0.07 ± 0.04; Fig. 2e). The FK506 treatment signature was also compared with those of more than 40 other deletion mutant strains or drug-treatments thought to affect

unrelated pathways, and none had statistically significant correlations. These data establish that genetic disruption of calcineurin function provides a close and specific phenocopy of treatment with FK506 or CsA.

To avoid generalizing from a single example, we also compared the effects of treatment of wild-type cells with 3-aminotriazole (3-AT) with the effects of deletion of the *HIS3* gene. *HIS3* encodes imidazoleglycerol phosphate dehydratase, which catalyzes the seventh step of the histidine biosynthetic pathway in yeast[22]; 3-AT is a competitive inhibitor of this enzyme that triggers a large transcriptional amino-acid starvation response[23]. Microarray analysis of wild-type and isogenic *his3*-deficient strains demonstrated the expected large genome-wide transcriptional responses (involving more than 1,000 ORFs) resulting from treatment with 3-AT (Fig. 3a) or from *HIS3* deletion (Fig. 3c). Quantitative comparison of the 3-AT treatment signature and the *his3* mutant signature showed a high level of correlation (ρ= 0.76 ± 0.02) that even extended to genes that experienced small changes in expression level (Fig. 3b). As a negative control, the correlations between the 3-AT treatment signature or the *his3* mutant signature and the calcineurin mutant strain were not statistically significant (ρ = 0.09 ± 0.06 and -0.01 ± 0.04, respectively). That both the calcineurin/FK506 and the *his3*/3-AT comparisons were highly correlated indicates that in many cases the expression profile resulting from a gene deletion closely resembles the expression profile of wild-type cells treated with an inhibitor of that gene's product.

'Decoder' strategy: Drug target validation with deletion mutants
Because pharmacological inhibition of different targets might give similar or identical expression profiles, simple comparison of drug signatures to mutant signatures is unlikely to unambiguously identify a drug's target. To overcome this limitation, an additional 'decoder' step is used. We first compare the expression profile of wild-type drug-treated cells to the expression profiles from a panel of genetic mutant strains, using a correlation coefficient metric. Mutant strains whose expression profile is similar to that of drug-treated wild-type cells are selected and subjected to drug treatment, generating the drug signature in the mutant strain (that is, the mutant drug signature). If the mutated gene encodes a protein involved in a pathway affected by the drug, we expect the drug signature in mutant cells to be different (or absent, for an ideal drug) from the drug signature seen in wild-type cells.

Fig. 2 Expression profiles from FK506-treated wild-type (wt) cells and a calcineurin-disruption mutant strain share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from FK506 treatment (*a* and *b*) or from genetic disruption of genes encoding calcineurin (*c*). *a*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from mock-treated strain R563 and Cy3-labeled cDNA (green) from strain R563 treated with 1 µg/ml FK506. *b*, Enlarged view of the boxed area in *a*. Arrowheads indicate specific ORFs induced or repressed. *c*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from strain R563 and Cy3-labeled cDNA (green) from strain MCY300 (deleted for the *CNA1,CNA2* catalytic subunits of calcineurin). Arrows indicate specific ORFs induced or repressed. *d*, The $\log_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the $\log_{10}$ of the expression ratio in the calcineurin mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. *e*, The $\log_{10}$ of the expression ratio for each ORF derived from the FK506 treatment hybridizations is plotted versus the $\log_{10}$

of the expression ratio in the *yer071c* mutant hybridizations. No ORFs were induced or repressed in both experiments.

To illustrate this, we treated the *his3* mutant strain with 3-AT. The signature pattern of altered gene expression resulting from treatment of the mutant strain with 3-AT was much less complex than that of the 3-AT signature in wild-type cells (Fig. 4). This is seen simply by examining plots of mean intensity of the hybridization signal (which approximately reflects level of expression) versus the expression ratio for each ORF (Fig. 4). Genes that were expressed at higher or lower levels in 3-AT treated cells or in *his3* mutant cells are shown as red and green dots, respectively. We analyzed the 3-AT signature in wild-type (Fig. 4a) and *his3* mutant cells (Fig. 4c), as well as the *his3* mutant strain signature (Fig. 4b). Whereas histidine limitation induced more than 1,000 transcription-level changes in the wild-type strain, few or no transcript level changes were induced by treatment of the *his3*-deletion strain with 3-AT. This indicates that with the growth conditions used, essentially all of the effects of 3-AT depend on or are mediated through the HIS3 gene product.

Applying this approach to the calcineurin signaling pathway showed the specificity of the method. The calcineurin mutant strain and strains with deletions in the genes encoding the most abundant immunophilins in yeast[12] (*CPH1* and *FPR1*) were treated with either FK506 or CsA to determine the profiles

of altered gene expression resulting from drug treatment of the mutant cells (that is, mutant +/- drug). We compared the drug signatures in the mutants to the wild-type drug signature using the correlation coefficient metric (Table 1). Although the signature generated by treatment of wild-type cells with FK506 was highly correlated to the calcineurin mutant strain signature ($\rho = 0.75 \pm 0.03$), it bore no similarity to the profile after treatment of the calcineurin mutant strain with FK506 ($\rho = -0.01 \pm 0.07$). This indicates that FK506 was unable to elicit its normal transcriptional response in the calcineurin mutant strain. Likewise, treatment of the *fpr1* mutant strain with FK506 elicited an expression profile that was not correlated to the FK506 signature in the wild-type strain ($\rho = -0.23 \pm 0.07$), indicating that the *FPR1* gene product is likely to be involved in the pathway affected by FK506. The same was true for the *cna fpr1* mutant strain. In contrast, treatment of the *cph1* mutant strain with FK506 generated an expression profile highly correlated with the wild-type FK506 expression profile ($\rho = 0.79 \pm 0.03$), indicating the *cph1* mutation did not block the mode of action of FK506 and thus is not directly involved in the pathway affected by FK506. We tabulated the change in expression in response to FK506 in different mutant strains for all ORFs with expression ratios greater than 1.8 in FK506-treated cells or in the calcineurin mutant strain (Fig. 5a).The calcineurin mutant strain signature and the FK506 responses in wild-type and the *cph1* mutant strain are similar, and there are no transcript-level changes (seen in black) for treatment of the calcineurin, *fpr1* and *cna fpr1* mutant strains with FK506 (Fig. 5a).

Similar experiments and analyses with CsA provided further validation of this approach. The expression profile elicited by treatment of wild-type cells with CsA was highly corre-

Table 1 Signature correlation of expression ratios as a result of FK506 treatment in various mutant strains

	wild-type +/-FK506	cna +/-FK506	fpr1 +/-FK506	cna fpr1 +/-FK506	cph1 +/-FK506
wild-type +/- FK506	0.93 ± 0.04	-0.01 ± 0.07	-0.23 ± 0.07	0.12 ± 0.07	0.79 ± 0.03

Signature correlation shows the absence of the FK506 signature specifically in the calcineurin (*cna*) and *fpr1* (major FK506 binding protein) deletion mutants. *cna* represents the mutant with deletions of the catalytic subunits of calcineurin, *CNA1* and *CNA2*. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/- FK506 experiments.

Fig. 3 Expression profiles from a *his3* mutant strain and wild-type (wt) cells treated with 3-AT share a genome-wide correlation. DNA microarray analysis showing changes in gene expression resulting from 3-AT treatment (*a*) or from genetic disruption of the *HIS3* gene (*c*). *a*, Pseudo-color image of the results of simultaneous hybridization of

Cy5-labeled cDNA (red) from mock-treated wild-type strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. *b*, Plot of the log₁₀ of the expression ratio for each ORF derived from the 3-AT treatment hybridizations is plotted versus the log₁₀ of the expression ratio in the *his3* mutant hybridizations. ORFs that were induced or repressed in both experiments are shown as green and red dots, respectively. The correlation of expression ratios applies not only to genes with large expression ratios (for example, *CHA1* and *ARG1*), but also extends to genes with expression ratios less than 2 (for example, *ILV1* and *CPH1*). *ILV1* is induced 1.9-fold and 1.5-fold, and *CPH1* is downregulated 1.9-fold

and 1.7-fold, in cells treated with 3-AT and *his3* mutant cells, respectively. Two ORFs do not fall on the line x = y. The leftmost point is the *HIS3* data point, which is induced by 3-AT treatment but which is not absent from the *his3* mutant strain. The other point is *YOR203w*. Both data points are labeled *HIS3* because hybridization to *YOR203w* is most likely due to *HIS3* mRNA, as *YOR203w* overlaps the *HIS3* open reading frame. *c*, Pseudo-color image of the results of simultaneous hybridization of Cy5-labeled cDNA (red) from wild-type strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. Arrowheads indicate specific ORFs induced or repressed.

lated to the profile elicited by mutation of the calcineurin genes (ρ = 0.71 ± 0.04), but did not correlate with the expression profile resulting from treatment of the calcineurin mutant strain with CsA (ρ = −0.05 ± 0.07; Table 2), indicating that the genetic deletion of calcineurin interfered with the ability of CsA to elicit its normal transcriptional response. Likewise, the CsA signature was essentially absent in CsA-treated *cph1* mutant cells, and the expression profile of CsA-treated *cph1* mutant cells correlated poorly to that of CsA-treated wild-type cells (ρ = 0.18 ± 0.07). Thus, the *CPH1* gene product was required for the CsA response seen in wild-type cells. Conversely, treatment of *fpr1* mutant cells with CsA resulted in an expression pattern very similar to the profile of CsA-treated wild-type cells (ρ = 0.77 ± 0.03), indicating that *FPR1* was not necessary for the CsA-mediated effects. Analysis of individual ORFs affected by CsA and their expression ratios over the entire set of experiments confirmed that *CPH1* and the genes encoding calcineurin, but not

FPR1, are necessary for the wild-type CsA response (Fig. 5*b*). The observation that the profiles resulting from FK506 or CsA drug treatment are similar to that of the calcineurin deletion mutant strain might allow the prediction that calcineurin was involved in the pathway affected by these drugs. But because the expression profile of the *fpr1* mutant strain did not bear a strong similarity to the wild-type drug expression profile for FK506, it is obvious that the drug treatment of the mutant strains was necessary to identify Fpr1, but not Cph1, as a potential FK506 drug target. In the same way, the 'decoder' strategy was necessary to identify Cph1, but not Fpr1, as a potential drug target for CsA.

'Decoder' approach can identify secondary drug effects

For a drug that has a single biochemical target, the strategy outlined above may be useful in target validation. In many cases, however, a compound may affect multiple pathways and elicit a very complex signature. 'Decoding' such a complex signature

Fig. 4 Treatment of the *his3* mutant strain with 3-AT shows nearly complete loss of 3-AT signature. A plot of the log₁₀ of the mean intensity of hybridization for each ORF versus the log₁₀ of its expression ratio for each experiment is shown next to a pseudo-color image of a representative portion of the microarray. ORFs that are induced or repressed at the 95% confidence level are shown in green and red, respectively. *a*, Expression profile from treatment of the wild-type (wt) strain with 3-AT. Cy5-labeled cDNA (red) from mock-treated strain R491 and Cy3-labeled cDNA (green) from strain R491 treated with 10 mM 3-AT. *b*, Expression profile

from the *his3* deletion strain. Cy5-labeled cDNA (red) from strain R491 and Cy3-labeled cDNA (green) from strain R1226, deleted for the *HIS3* gene. *c*, Expression profile of treatment of the *his3* deletion strain with 3-AT. Cy3-labeled cDNA (red) from *his3*-deleted strain R1226 and Cy5-labeled cDNA (green) from strain R1226 treated with 10 mM 3-AT. Arrowheads indicate the DNA probe and data point corresponding to the *HIS3* gene. The blue dashed line represents the threshold below which errors tend to increase rapidly because spot intensities are not sufficiently above background intensity.

Table 2 Signature correlati n f expression ratios as a result of CsA treatment in various mutant strains					
wild-type +/– CsA	wild-type +/–CsA	cna +/–CsA	fpr1 +/–CsA	cna cph1 +/–CsA	cph1 +/–CsA
wild-type +/– CsA	0.94 ± 0.04	–0.05 ± .07	0.77 ± 0.03	–0.11 ± 0.07	0.18 ± 0.07

Signature correlation shows the absence of the CsA signature specifically in the calcineurin (cna) and cph1 (cyclophilin) deletion mutants. cna represents the mutant with deletions of the catalytic subunits of calcineurin, CNA1 and CNA2. The correlation coefficient reported in the first column represents the correlation between two pairs of hybridizations from independent wild-type +/– CsA experiments.

into the effects mediated through the intended target (the 'on-target signature') and those mediated through unintended targets (the 'off-target' signature) might be useful in evaluating a compound's specificity. Our 'decoder' strategy is based on the premise that 'off-target' signature should be insensitive to the genetic disruption of the primary target.

To determine whether the 'decoder' approach could identify an 'off-target' profile, we looked for a drug-responsive gene whose expression is insensitive to deletion of the primary target. To increase the likelihood of observing such genes, the same strains described in Tables 1 and 2 were treated with higher concentrations (50 µg/ml) of FK506. This led to a much more complex expression profile in wild-type cells, indicating that at this higher concentration, FK506 was inhibiting or activating additional targets. Several of the ORFs in this expanded FK506-induced expression profile were not affected by the calcineurin, cph1 or fpr1 mutations, as drug treatment of these mutant strains did not block their presence in the FK506 expression signature (Fig. 6). This indicates that FK506 was triggering changes in transcript levels of many genes through pathways independent of calcineurin, CPH1 and FPR1. Many of the upregulated ORFs in the 'off-target' pathway were genes reported to be regulated by the transcriptional activator Gcn4 (ref. 24). In some strains, a reporter gene under GCN4 control was induced in response to FK506 treatment[25]. To determine whether GCN4 is involved in this pathway that is independent of calcineurin, CPH1 and FPR1, we analyzed the effects of treatment with high-dose FK506 on global gene expression in a strain with a GCN4 deletion (Fig. 6). Of the 41 ORFs with calcineurin-independent expression ratios greater than 4, 32 were not induced in the gcn4 mutant, indicating that their induction by FK506 was GCN4-dependent. Not all GCN4-regulated genes were induced by FK506. This FK506-induced subset of GCN4-regulated genes may be those most sensitive to subtle changes in Gcn4 levels, or perhaps other regulatory circuits prevent FK506 activation of some GCN4-regulated genes. Seven of the remaining nine ORFs induced by FK506 were independent of

both the calcineurin and GCN4 pathways. The simplest explanation is that FK506 inhibits or activates additional pathways. Members of this class include SNQ2 and PDR5, genes that enc de drug efflux pumps with structural homology to mammalian multiple drug resistance proteins[26]. FK506 may interact directly with Pdr5 to inhibit its function[27]. Our results indicate that treatment with FK506 leads to four-fold-to-sixfold induction of PDR5 mRNA levels. YOR1, another gene that can confer drug resistance, is also induced threefold-to-fourfold by FK506. Thus, drug treatment of strains with mutations in the primary targets can prove useful in identifying effects mediated by secondary drug targets, including the nature and extent of newly discovered and previously unsuspected pathways affected by the drug.

We describe here a method for drug target validation and the identification of secondary drug target effects that uses DNA microarrays to survey the effects of drugs on global gene expression patterns. We established that genetic and pharmacologic inhibition of gene function can result in extremely similar changes in gene expression. We also demonstrated that one can confirm a potential drug target by treating a deletion mutant defective in the gene encoding the putative target. Drug-mediated signatures from strains with mutations in pathways or processes directly or indirectly affected by the drug bore little or

Fig. 5 Response of FK506 and CsA signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 1.8 in response to treatment with 1 µg/ml FK506 (a) or 50 µg/ml CsA (b) are listed (left side) and their expression ratios in the indicated strain are shown on the green (induction)-red (repression) color scale. a, Calcineurin (cna) mutant and FK506 treatment signature genes are in the first two columns. Almost all FK506 signature genes have expression ratios near unity in deletion strains involved in pathways affected by FK506 (calcineurin, fpr1 and cna fpr1 mutants) but not in deletion strains in unrelated pathways (cph1). b, Calcineurin (cna) mutant and CsA treatment signature genes are in the first two columns. Almost all CsA signature genes have expression ratios near unity in deletion strains involved in pathways affected by CsA (calcineurin, cph1 and cna cph1 mutants) but not in deletion strains in unrelated pathways (fpr1).

Strain: wt cna fpr1 cph1 cna fpr1 gen4
FK506: + + + + + +

Fold repression   Fold induction
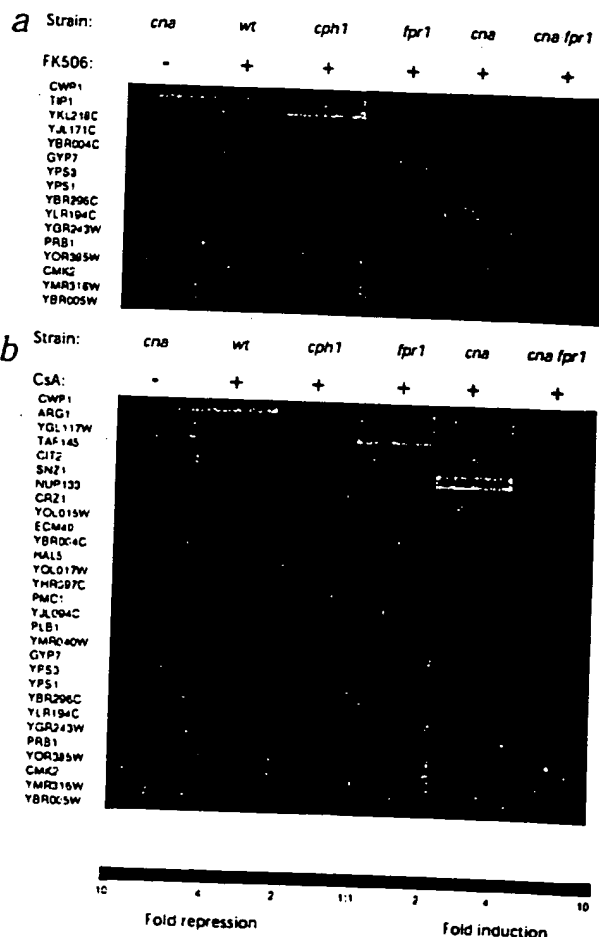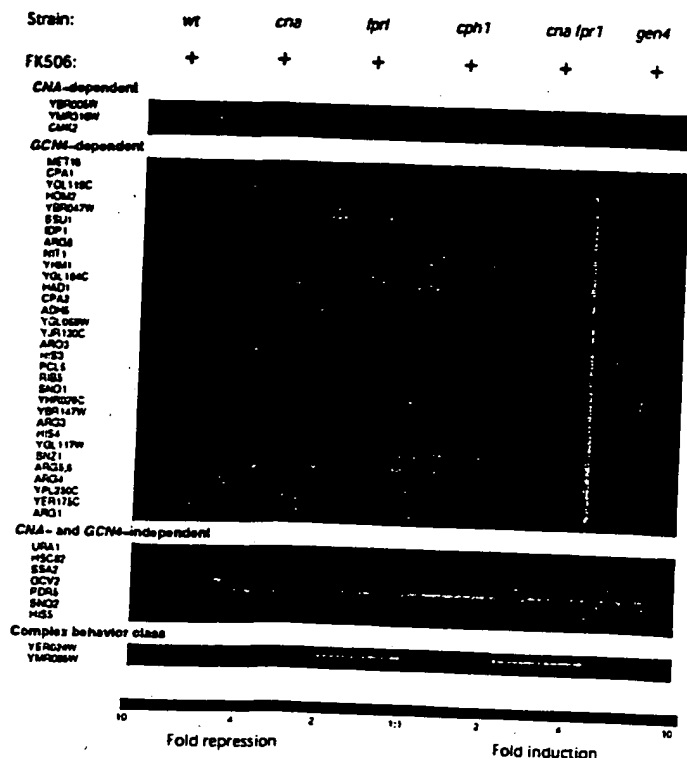
Fig. 6   Response of FK506 signature genes in strains with deletions in different genes. Genes with expression ratios greater than a factor of 4 in at least one experiment are listed and their expression ratios in the indicated strain are shown in the green (induction)–red (repression) color scale. The genes have been divided into classes corresponding to these expected behaviors: 'CNA-dependent' genes respond to FK506 (50 µg/ml) except when either calcineurin genes or FPR1 or both are deleted: 'GCN4-dependent' genes respond to FK506 except when GCN4 is deleted. These genes still respond to FK506 when calcineurin genes or FPR1 or CPH1 are deleted; that is, their responses are not mediated by calcineurin, Cph1, or Fpr1. 'CNA- and GCN4-independent' genes respond to FK506 in all deletion strains tested. A 'complex behavior' class is provided for those genes that did not match the model of FK506 response mediated through calcineurin or Fpr1 or separately through Gcn4.

penile erection. It is possible that application of the 'decoder' to other compounds may show that they too have a potent activity against a target distinct from their intended target.

The ability to decode drug effects is dependent on the availability of functionally 'targetless' cells. In yeast, this is being achieved by systematically disrupting each yeast gene (Saccharomyces Deletion Consortium: http://sequence-www.stanford.edu/group/yeast_deletion_project/deletion.html). Efforts are underway to obtain expression profiles from each deletion mutant strain. Determining signatures resulting from inactivation of essential genes presents a unique problem, but it may be possible to do so by examining heterozygotes or by using a controllable promoter to reduce expression of the essential gene. Although it is already feasible to test several compounds in dozens of yeast strains, another challenge for the 'decoder' strategy will be the efficient selection of the mutants with deletions in genes most likely to encode the intended drug target. The signature correlation plots described are one metric that could be used as part of that selection process, but others need to be explored. Applying the 'decoder' to mammalian cells presents additional challenges. It is considerably more difficult to isolate functionally 'targetless' cells. Strategies involving titratable promoters, known specific inhibitors, anti-sense RNAs, ribozymes, and methods of targeting specific proteins for degradation are possible and should be tested. Another limitation is that not all cell types express the same set of genes and therefore 'off-target' effects may be different in different cell types. In addition, applying the 'decoder' to human cells will also require technical improvements that allow expression profiling from a small number of cells. Even the broader question of whether the insensitivity of 'off-target' signatures to the disruption of the main target is the exception or the rule can only be answered by the accumulation of more data. Barkai and Leibler, however, have argued in favor of robustness of biological networks, indicating that drug perturbations ('off-target' signatures) may be robust even when the system is subjected to another perturbation (such as a genetic disruption)(ref. 28). Many practical developments will be necessary if the 'decoder' concept is to be broadly applied.

no similarity to the wild-type drug expression profile. In contrast, drug-mediated signatures from strains with mutations in genes involved in pathways unrelated to the drug's action showed extensive similarity to the wild-type drug signature. By applying this approach to a drug that affects multiple pathways (FK506), we were able to decode a complex signature into component parts, including the identification of an 'off-target' signature that was mediated through pathways independent of calcineurin or the Fpr1 immunophilin.

Discussion

It is well-established that high-throughput biochemical screening can identify potent inhibitory compounds against a given target. The 'decoder' approach described here complements this process by evaluating the equally important property of specificity: the tendency of a compound to inhibit pathways other than that of its intended target. The ability to observe such 'off-target' effects will likely be useful in several ways. Profiling compounds with known toxicities will allow the development of a database of expression changes associated with particular toxicities. Recognition of potential toxicities in the 'off-target' signatures of otherwise promising compounds then may allow earlier identification of those likely to fail in clinical trials. Comparing the extent and peculiarities of 'off-target' signatures of promising drug candidates could provide a new way to group compounds by their effects on secondary pathways, even before those effects are understood. This may prove to be an alternative, potentially more effective, way to select compounds for animal and clinical trials. Some drugs are more effective against a related protein than against the originally intended target. Sildenafil (Viagra™), for example, was initially developed as a phosphodiesterase inhibitor to control cardiac contractility, but was found to be highly specific for phosphodiesterase 5, an isozyme whose inhibition overcomes defects in

Expression arrays have been used mainly as an initial screen for genes induced in a particular tissue or process of interest by focusing on genes with large expression ratios. We have found, however, that effort to refine experimental protocols and repeat experiments increases the reliability of the data and permits new applications. For example, it provides a larger set

1298

Table 3	Yeast strains used	
Strain	Relevant genotype	Reference
YPH499	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1	(34)
R563	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3	(this study)
R558	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 fpr1::HIS3	(this study)
R567	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cph1::HIS3	(this study)
MCY300	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3	(21)
R132	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 cph1::kanr	(this study)
R133	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 cna1Δ1::hisG cna2Δ1::HIS3 fpr1::kanr	(this study)
R559	Mata ura3-52 lys2-801 ade2-101 trp1-Δ63 his3-Δ200 leu2-Δ1 his3::HIS3 gcn4::LEU2	(this study)
BY4719	Mata trp1-Δ63 ura3-Δ0	(35)
BY4738	Matα trp1-Δ 63 ura3-Δ0	(35)
R491	Mata/α BY4719 X BY4738	(35)
BY4728	Mata his3-Δ200 trp1-Δ63 ura3-Δ0	(this study)
BY4729	Matα his3-Δ 200 trp1-Δ63 ura3-Δ0	(35)
R1226	Mata/α BY4728 X BY4729	(35) (this study)

of genes at higher confidence levels that serve as a more unique signature for a given protein perturbation. In addition, it allows subtle signatures to be detected, when, for example, a protein is only partially inhibited. This may enable clinical monitoring of small changes in protein function in disease or toxicity states before they could otherwise be detected. Because the functions of many genes detected on transcript arrays are known, these microarrays are powerful tools that provide detailed information about a cell's physiology. For example, changes in the flux through a metabolic pathway are reflected in transcriptional changes in genes in the pathway[7]. Furthermore, it may be possible to indirectly measure protein activity levels from expression profiling data (S.F., et al., unpublished data). Thus, although the eventual development of genomic methods allowing the direct measurement of all cellular protein levels will be an important achievement, transcript array technology offers an immediate and robust means of evaluating the effects of various treatments on gene expression and protein function.

Methods

Construction, growth and drug treatment of yeast strains. The strains used in this study (Table 3) were constructed by standard techniques[29]. To construct strain R559, strain R563 was transformed to Leu⁺ with plasmid pM12 digested by SalI and MluI (provided by A. Hinnebusch and T. Dever). Strains R132 and R133 were constructed by transforming the bacterial kanamycin resistance cassette[30] flanked by genomic DNA from the CPH1 and FPR1 loci, respectively, and selecting for G418-resistant colonies. For experiments with FK506, cells were grown for three generations to a density of 1×10^7 cells/ml in YAPD medium (YPD plus 0.004% adenine) supplemented with 10 mM calcium chloride as described[31]. Where indicated, FK506 was added to a final concentration of 1 µg/ml 0.5 h after inoculation of the culture or to 50 µg/ml 1 h before cells were collected. CsA was used at a final concentration of 50 µg/ml. Cells were broken by standard procedures[32] with the following modifications: Cell pellets were resuspended in breaking buffer (0.2 M Tris HCl pH 7.6, 0.5 M NaCl, 10 mM EDTA, 1% SDS), vortexed for 2 min on a VWR multi-tube vortexer at setting 8 in the presence of 60% glass beads (425–600 µm mesh; Sigma) and phenol:chloroform (50:50, volume/volume). After separation of the phases, the aqueous phase was re-extracted and ethanol-precipitated. Poly A⁺ RNA was isolated by two sequential chromatographic purifications over oligo dT cellulose (New England Biolabs, Beverly, Massachusetts) using established protocols[32].

For experiments using 3-AT, wild-type or his3/his3 cells were grown to early logarithmic phase in SC medium, pelleted and resuspended in SC medium lacking histidine for 1 hr in the presence or absence of 10 mM 3-

AT, as indicated. Cells were harvested and mRNA isolated as above. FK506 was obtained from the Swedish Hospital Pharmacy (Seattle, Washington) and purified to homogeneity by ethyl acetate extraction by J. Simon (Fred Hutchinson Cancer Research Center, Seattle, Washington). CsA was obtained from Alexis Biochemicals (San Diego, California); 3-AT was from Sigma.

Preparation and hybridization of the labeled sample. Fluorescently-labeled cDNA was prepared, purified and hybridized essentially as described[7]. Cy3- or Cy5-dUTP (Amersham) was incorporated into cDNA during reverse transcription (Superscript II; Life Technologies) and purified by concentrating to less than 10 µl using Microcon-30 microconcentrators (Amicon, Houston, Texas). Paired cDNAs were resuspended in 20–26 µl hybridization solution (3 × SSC, 0.75 µg/ml polyA DNA, 0.2% SDS) and applied to the microarray under a 22- x 30-mm coverslip for 6 h at 63 °C, all according to a published method[7].

Fabrication and scanning of microarrays. PCR products containing common 5′ and 3′ sequences (Research Genetics, Huntsville, Alabama) were used as templates with amino-modified forward primer and unmodified reverse primers to PCR amplify 6,065 ORFs from the S. cerevisiae genome. Our first-pass success rate was 94%. Amplification reactions that gave products of unexpected sizes were excluded from subsequent analysis. ORFs that could not be amplified from purchased templates were amplified from genomic DNA. DNA samples from 100-µl reactions were isopropanol-precipitated, resuspended in water, brought to a final concentration of 3× SSC in a total volume of 15 µl, and transferred to 384-well microtiter plates (Genetix Limited, Christchurch, Dorset, England). PCR products were spotted onto 1 x 3-inch polylysine-treated glass slides by a robot built essentially according to defined specifications[3,6,7] (http://cmgm.stanford.edu/pbrown/MGuide). After being printed, slides were processed according to published protocols[7].

Microarrays were imaged on a prototype multi-frame CCD camera in development at Applied Precision (Issaquah, Washington). Each CCD image frame was approximately 2-mm square. Exposure times of 2 s in the Cy5 channel (white light through Chroma 618–648 nm excitation filter, Chroma 657–727 nm emission filter) and 1 s in the Cy3 channel (Chroma 535–560 nm excitation filter, Chroma 570–620 nm emission filter) were done consecutively in each frame before moving to the next, spatially contiguous frame. Color isolation between the Cy3 and Cy5 channels was about 100:1 or better. Frames were 'knitted' together in software to make the complete images. The intensity of spots (about 100 µm) were quantified from the 10-µm pixels by frame-by-frame background subtraction and intensity averaging in each channel. Dynamic range of the resulting spot intensities was typically a ratio of 1,000 between the brightest spots and the background-subtracted additive error level. Normalization between the channels was accomplished by normalizing each channel to the mean intensities of all genes. This procedure is nearly equivalent to normalization between channels using the intensity

ratio of genomic DNA spots', but is possibly more robust, as it is based on the intensities of several thousand spots distributed over the array.

Signature correlation coefficients and their confidence limits. Correlation coefficients between the signature ORFs of various experiments were calculated using:

$$\rho = \sum_k x_k y_k / (\sum_k x_k^2 \sum_k y_k^2)^{1/2}$$

where x_k is the $\log_{10}$ of the expression ratio for the kth gene in the x signature, and y_k is the $\log_{10}$ of the expression ratio for the kth gene in the y signature. The summation is over those genes that were either up- or down-regulated in either experiment at the 95% confidence level. These genes each had a less than 5% chance of being actually unregulated (having expression ratios departing from unity due to measurement errors alone). This confidence level was assigned based on an error model which assigns a lognormal probability distribution to each gene's expression ratio with characteristic width based on the observed scatter in its repeated measurements (repeated arrays at the same nominal experimental conditions) and on the individual array hybridization quality. This latter dependence was derived from control experiments in which both Cy3 and Cy5 samples were derived from the same RNA sample. For large numbers of repeated measurements the error reduces to the observed scatter. For a single measurement the error is based on the array quality and the spot intensity.

Random measurement errors in the x and y signatures tend to bias the correlation towards zero. In most experiments, most genes are not significantly affected but do show small random measurement errors. Selecting only the '95% confidence' genes for the correlation calculation, rather than the entire genome, reduces this bias and makes the actual biological correlations more apparent.

Correlations between a profile and itself are unity by definition. Error limits on the correlation are 95% confidence limits based on the individual measurement error bars, and assuming uncorrelated errors[33]. They do not include the bias mentioned above; thus, a departure of ρ from unity does not necessarily mean that the underlying biological correlation is imperfect. However, a correlation of 0.7 ± 0.1, for example, is very significantly different from zero. Small (magnitude of $\rho < 0.2$) but formally significant correlation in the tables and text probably are due to small systematic biases in the Cy5/Cy3 ratios that violate the assumption of independent measurement errors used to generate the 95% confidence limits. Therefore, these small correlation values should be treated as not significant. A likely source of uncorrected systematic bias is the partially corrected scanner detector nonlinearity that differently affects the Cy3 and Cy5 detection channels.

The 1 µg/ml FK506 treatment signature was compared with more than 40 unrelated deletion mutant strain or drug signatures. These control profiles had correlation coefficients with the FK506 profile that were distributed around zero (mean $\rho = -0.03$) with a standard deviation of 0.16 (data not shown), and none had correlations greater than $\rho = 0.38$. Similarly, the calcineurin mutant strain signature correlated well with the CsA treatment signature ($\rho = 0.71 \pm 0.04$) but not with the signatures from the negative controls (mean $\rho = -0.02$ with a standard deviation of 0.18).

Quality controls. End-to-end checks on expression ratio measurement accuracy were provided by analyzing the variance in repeated hybridizations using the same mRNA labeled with both Cy3 and Cy5, and also using Cy3 and Cy5 mRNA samples isolated from independent cultures of the same nominal strain and conditions. Biases undetected with this procedure, such as gene-specific biases presumably due to differential incorporation of Cy3- and Cy5-dUTP into cDNA, were minimized by doing hybridizations in fluor-reversed pairs, in which the Cy3/Cy5 labeling of the biological conditions was reversed in one experiment with respect to the other. The expression ratio for each gene is then the ratio of ratios between the two experiments in the pair. Other biases are removed by algorithmic numerical de-trending. The magnitude of these biases in the absence of de-trending and fluor reversal is typically about 30% in the ratio, but may be as high as twofold for some ORFs.

Expression ratios are based on mean intensities over each spot. Some

smaller spots have fewer image pixels in the average. This does not degrade accuracy noticeably until the number of pixels falls below ten. In which case the spot is rejected from the data set. 'Wander' of spot positions with respect to the nominal grid is adaptively tracked in array subregions by the image processing software. Unequal spot 'wander' within a subregion greater than half-a-spot spacing is a difficulty for the automated quantitating algorithms; in this case, the spot is rejected from analysis based on human inspection of the 'wander'. Any spots partially overlapping are excluded from the data set. Less than 1% of spots typically are rejected for these reasons.

1. Schena, M., Shalon, D., Davis, R.W. & Brown, P.O. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* 270, 467–470 (1995).
2. Schena, M. *et al.* Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. USA* 93, 10614–10619 (1996).
3. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645 (1996).
4. Lockhart, D.J. *et al.* Expression monitoring by hybridization to high-density oligonucleotide arrays. *Nature Biotechnol.* 14, 1675–1680 (1996).
5. DeRisi, J. *et al.* Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nature Genet.* 14, 457–460 (1996).
6. Heller, R.A. *et al.* Discovery and analysis of inflammatory disease-related genes using cDNA microarrays. *Proc. Natl. Acad. Sci. USA* 94, 2150–2155 (1997).
7. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686 (1997).
8. Lashkari, D.A. *et al.* Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* 94, 13057–13062 (1997).
9. Wodicka, L., Dong H., Mittman, M. Ho, M.-H. & Lockhart, D.J., Genome-wide expression monitoring in Saccharomyces cerevisiae. *Nature Biotechnol.* 15, 1359–1367 (1997).
10. Cho, R.J. *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73 (1998).
11. Gray, N.S. *et al.* Exploiting chemical libraries, structure, and genomics in the search for kinase inhibitors. *Science* 281, 533–538 (1998).
12. Cardenas, M.E., Lorenz, M., Hemenway, C. & Heitman, J. Yeast as model T cells. *Perspect. Drug Discovery Design* 2, 103–126 (1994).
13. Klee, C.B., Ren, H. & Wang, X. Regulation of the calmodulin-stimulated protein phosphatase, calcineurin. *J. Biol. Chem.* 273, 13367–13370 (1998).
14. Tanida, I., Hasegawa, A., Iida, H., Ohya, Y. & Anraku, Y. Cooperation of calcineurin and vacuolar H(+)-ATPase in intracellular Ca2+ homeostasis of yeast cells. *J. Biol. Chem.* 270, 10113–10119 (1995).
15. Moser, M.J., Geiser, J.R. & Davis, T.N. Ca2+-calmodulin promotes survival of pheromone-induced growth arrest by activation of calcineurin and Ca2+-calmodulin-dependent protein kinase. *Mol. Cell. Biol.* 16, 4824–4831 (1996).
16. Mizunuma, M., Hirata, D., Miyahara, K., Tsuchiya, E. & Miyakawa, T. Role of calcineurin and Mpk1 in regulating the onset of mitosis in budding yeast. *Nature* 392, 303–306 (1998).
17. Yazdanbakhsh, K., Choi, J.W., Li, Y., Lau, L.F. & Choi, Y. Cyclosporin A blocks apoptosis by inhibiting the DNA binding activity of the transcription factor Nur77. *Proc. Natl. Acad. Sci. USA* 92, 437–441 (1995).
18. Molkentin, J.D. *et al.* A calcineurin-dependent transcriptional pathway for cardiac hypertrophy. *Cell* 93, 215–228 (1998)
19. Mansuy, I.M., Mayford, M., Jacob, B., Kandel, E.R. & Bach, M.E. Restricted and regulated overexpression reveals calcineurin as a key component in the transition from short-term to long-term memory. *Cell* 92, 39–49 (1998).
20. Schreiber, S.L. & Crabtree, G.R. The mechanism of action of cyclosporin A and FK506. *Immunol. Today* 13, 136–142 (1992).
21. Cyert, M.S., Kunisawa, R., Kaim, D. & Thorner, J. Yeast has homologs (CNA1 and CNA2 gene products) of mammalian calcineurin, a calmodulin-regulated phosphoprotein phosphatase. *Proc. Natl. Acad. Sci. USA* 88, 7376–7380 (1991).
22. Jones, E.W. & Fink, G.R. in *The Molecular Biology of the Yeast Saccharomyces: Metabolism and Gene Expression* (eds. Strathern, J.N., Jones, E.W. & Broach, J.R.) 181–299 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1982).
23. Hinnebusch, A. Translational regulation of yeast GCN4. *J. Biol. Chem.* 272, 21661–21664 (1997).
24. Hinnebusch, A.G. in *The Molecular and Cellular Biology of the Yeast*

Saccharomyces: Gene Expression. (eds. Jones, E.W., Pringle, J.R. & Broach, J.R.) 319–414 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1992).

25. Heitman, J. *et al.* The immunosuppressant FK506 inhibits amino acid import in *Saccharomyces cerevisiae. Mol. Cell. Biol.* 13, 5010–5019 (1993).

26. Balzi, E. & Goffeau, A. Yeast multidrug resistance: the PDR network. *J. Bioenerg. Biomembr.* 27, 71–76 (1995).

27. Egner, R., Rosenthal, F.E., Kralli, A., Sanglard, D. & Kuchler, K. Genetic separation of FK506 susceptibility and drug transport in the yeast Pdr5 ATP-binding cassette multidrug resistance transporter *Mol. Biol. Cell* 9, 523–543 (1998).

28. Barkal, N. & Leiber, S. Robustness in simple biochemical networks. *Nature* 387, 913–917 (1997).

29. Schiestl, R.H., Manivasakam, P., Woods, R.A. & Gietz, R.D. Introducing DNA into yeast by transformation. *Methods: A companion to Methods in Enzymology* 5, 79–85 (1993).

30. Wach, A., Brachat, A., Pohlmann, R. & Philippsen, P. New heterologous mod-ules for classical or PCR-based gene disruptions in *Saccharomyces cerevisiae. Yeast* 10, 1793–1808 (1994).

31. Garrett-Engele, P., Mollanen, B. & Cyert, M.S. Calcineurin, the Ca2+/calmod-ulin-dependent protein phosphatase, is essential in yeast mutants with cell in-tegrity defects and in mutants that lack a functional vacuolar H(+)-ATPase. *Mol. Cell. Biol.* 15, 4103–4114 (1995).

32. Ausubel, F.M. *et al.* in *Current Protocols in Molecular Biology* 13.12.1–13.12.5 (eds. Ausubel, F.M., *et al.*)(John Wiley & Sons, New York, 1993).

33. Bulmer, M.G. in *Principles of Statistics* 224–225 (Dover Publications, New York, 1979).

34. Sikorski, R.S. & Hieter, P. A system of shuttle vectors and yeast host strains des-ignated for efficient manipulation of DNA in *Saccharomyces cerevisiae. Genetics* 122, 19–27 (1989).

35. Brachmann, C.B. *et al.* Designer deletion strains derived from Saccharomyces cerevisiae S288C: A useful set of strains and plasmids for PCR-mediated gene disruption and other applications. *Yeast* 14, 115–132 (1998).

co mosaic viral RNA was obtained by phenol and chloroform extractions of the virus and precipitated from ethanol. CA-NC assembly reactions in the presence of noncognate RNAs were identical to those given in (9). In the absence of RNA, CA-NC cones formed under the following conditions: 300 µM CA-NC, 1 M NaCl, and 50 mM tris-HCl (pH 8.0) at 37°C for 60 min. In the absence of exogenous RNA, neither cones nor cylinders formed at concentrations of 0.5 M NaCl or below. Absorption spectra demonstrated that our CA-NC preparations were not contaminated with *Escherichia coli* RNA (estimated lower detection limit was ~1 base/protein molecule). To control for even lower levels of RNA contamination, we preincubated the CA-NC protein with 0.5 mg/ml ribonuclease A (Type 1-AS, S4 Kunitz U/mg, Sigma) for 1 hour at 4°C, which then formed cones normally.

13. V. Y. Klishko, data not shown.
14. M. Ge and K. Sattler, *Chem. Phys. Lett.* 220, 192 (1994).
15. A. Krishnan *et al.*, *Nature* 388, 451 (1997).
16. L. B. Kong *et al.*, *J. Virol.* 72, 4403 (1998).
17. Assembly mixtures were deposited on holey carbon grids, blotted briefly with filter paper, plunged into liquid ethane, and transferred to liquid nitrogen. Frozen grids were transferred to a Philips 420 TEM equipped with a Gatan cold stage system, and images of particles in vitreous ice were recorded under low dose conditions at 36,000× magnification and ~1.6-µm defocus.
18. J. T. Finch, data not shown.
19. R. A. Crowther, *Proceedings of the Third John Innes Symposium* (1976), pp. 15–25; E. Kellenberger, M. Häner, M. Wurtz, *Ultramicroscopy* 9, 139 (1982); J. Seymore and D. J. DeRosier, *J. Microsc.* 148, 195 (1987).
20. M. V. Nermut, C. Grief, S. Hashmi, D. J. Hockley, *AIDS Res. Hum. Retroviruses* 9, 929 (1993); M. V. Nermut *et al.*, *Virology* 198, 288 (1994); E. Barklis, J. McDermott, S. Wilkens, S. Fuller, D. Thompson, *J. Biol. Chem.* 273, 7177 (1998); E. Barklis *et al.*, *EMBO J.* 16, 1199 (1997); M. Yeager, E. M. Wilson-Kubalek, S. G. Weiner, P. O. Brown, A. Rein, *Proc. Natl. Acad. Sci. U.S.A.* 95, 7299 (1998).
21. J. T. Finch *et al.*, unpublished observations.
22. V. M. Vogt, in (2), pp. 27–70.
23. M. A. McClure, M. S. Johnson, D.-F. Feng, R. F. Doolittle, *Proc. Natl. Acad. Sci. U.S.A.* 85, 2469-2473 (1988).
24. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
25. We thank C. Hill for very helpful discussions on the relationship between viral cores and fullerene cones, D. Hobbs for refining the ChemDraw3D images of cones, G. Stubbs for a gift of tobacco mosiac virus, J. McCutcheon for the plasmid used to prepare ribosomal RNA, and K. Albertine and N. Chandler of the University of Utah Shared Electron Microscopy facility for their support and encouragement. Supported by grants from NIH and from the Huntsman Cancer Institute (to W.I.S.).

The Transcriptional Program in the Response of Human Fibroblasts to Serum

Vishwanath R. Iyer, Michael B. Eisen, Douglas T. Ross, Greg Schuler, Troy Moore, Jeffrey C. F. Lee, Jeffrey M. Trent, Louis M. Staudt, James Hudson Jr., Mark S. Boguski, Deval Lashkari, Dari Shalon, David Botstein, Patrick O. Brown*

The temporal program of gene expression during a model physiological response of human cells, the response of fibroblasts to serum, was explored with a complementary DNA microarray representing about 8600 different human genes. Genes could be clustered into groups on the basis of their temporal patterns of expression in this program. Many features of the transcriptional program appeared to be related to the physiology of wound repair, suggesting that fibroblasts play a larger and richer role in this complex multicellular response than had previously been appreciated.

The response of mammalian fibroblasts to serum has been used as a model for studying growth control and cell cycle progression (1). Normal human fibroblasts require growth factors for proliferation in culture; these growth factors are usually provided by fetal bovine serum (FBS). In the absence of growth factors, fibroblasts enter a nondividing state, termed G_0, characterized by low metabolic activity. Addition of FBS or purified growth factors induces proliferation of the fibroblasts; the changes in gene expression that accompany this proliferative response have been the subject of many studies, and the responses of dozens of genes to serum have been characterized.

We took a fresh look at the response of human fibroblasts to serum, using cDNA microarrays representing about 8600 distinct human genes to observe the temporal program of transcription that underlies this response. Primary cultured fibroblasts from human neonatal foreskin were induced to enter a quiescent state by serum deprivation for 48 hours and then stimulated by addition of medium containing 10% FBS (2). DNA microarray hybridization was used to measure the temporal changes in mRNA levels of 8613 human genes (3) at 12 times, ranging from 15 min to 24 hours after serum stimulation. The cDNA made from purified mRNA from each sample was labeled with the fluorescent dye Cy5 and mixed with a common reference probe consisting of cDNA made from purified mRNA from the quiescent

V. R. Iyer and D. T. Ross, Department of Biochemistry, Stanford University School of Medicine, Stanford CA 94305, USA. M. B. Eisen and D. Botstein, Department of Genetics, Stanford University School of Medicine, Stanford CA 94305, USA. G. Schuler and M. S. Boguski, National Center for Biotechnology Information, Bethesda MD 20894, USA. T. Moore and J. Hudson Jr., Research Genetics, Huntsville, AL 35801, USA. J. C. F. Lee, D. Lashkari, D. Shalon, Incyte Pharmaceuticals, Fremont, CA 94555, USA. J. M. Trent, Laboratory of Cancer Genetics, National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892, USA. L. M. Staudt, Metabolism Branch, Division of Clinical Sciences, National Cancer Institute, Bethesda, MD 20892, USA. P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford CA 94305, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

Fig. 1. The same section of the microarray is shown for three independent hybridizations comparing RNA isolated at the 8-hour time point after serum treatment to RNA from serum-deprived cells. Each microarray contained 9996 elements, including 9804 human cDNAs, representing 8613 different genes. mRNA from serum-deprived cells was used to prepare cDNA labeled with Cy3-deoxyuridine triphosphate (dUTP), and mRNA harvested from cells at different times after serum stimulation was used to prepare cDNA labeled with Cy5-dUTP. The two cDNA probes were mixed and simultaneously hybridized to the microarray. The image of the subsequent scan shows genes whose mRNAs are more abundant in the serum-deprived fibroblasts (that is, suppressed by serum treatment) as green spots and genes whose mRNAs are more abundant in the serum-treated fibroblasts as red spots. Yellow spots represent genes whose expression does not vary substantially between the two samples. The arrows indicate the spots representing the following genes: 1, protein disulfide isomerase-related protein P5; 2, IL-8 precursor; 3, EST AA057170; and 4, vascular endothelial growth factor.

culture (time zero) labeled with a second fluorescent dye, Cy3 (4). The color images of the hybridization results (Fig. 1) were made by representing the Cy3 fluorescent image as green and the Cy5 fluorescent image as red and merging the two color images.
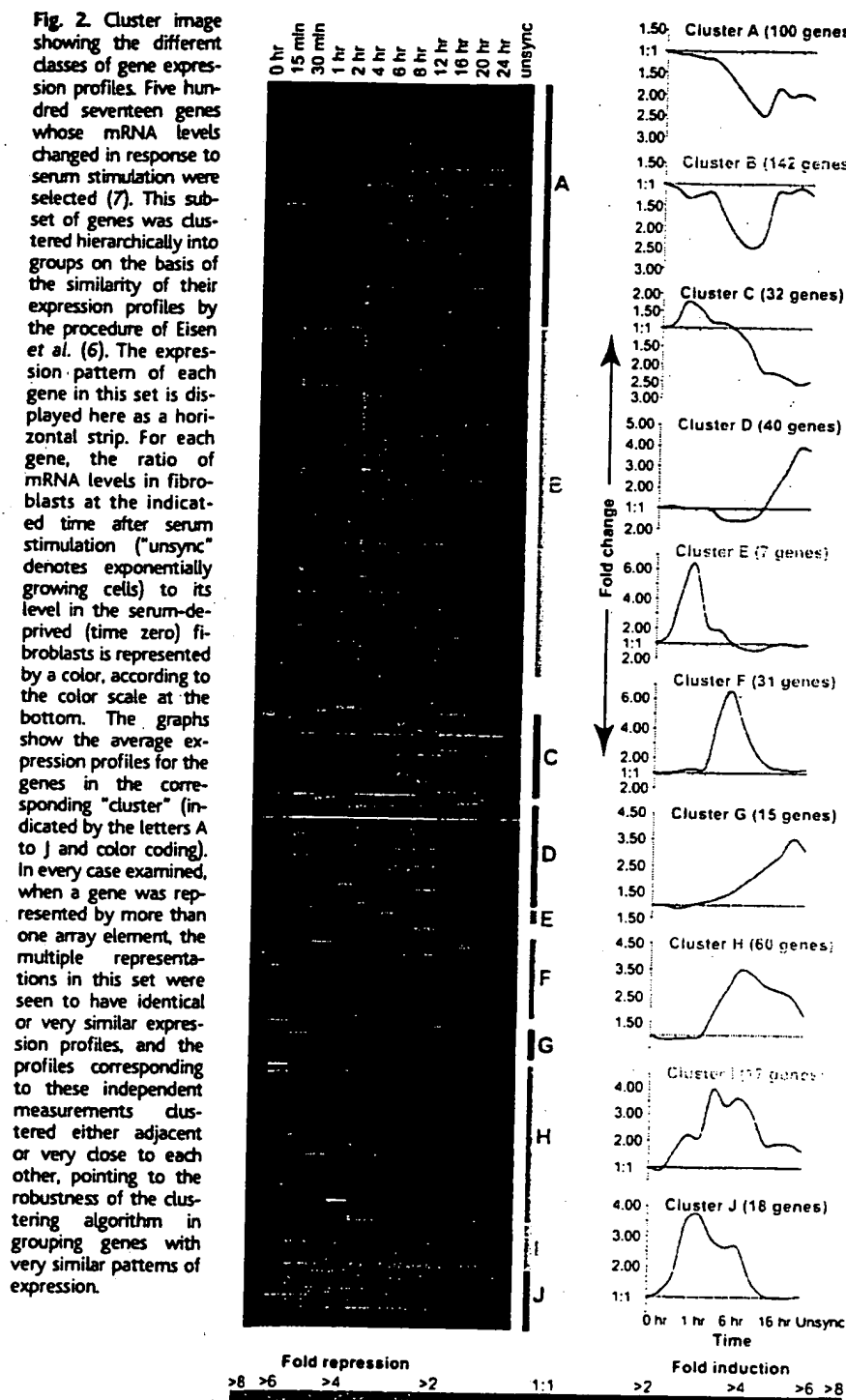
Diverse temporal profiles of gene expression could be seen among the 8613 genes surveyed in this experiment (Fig. 2); many of these genes (about half) were unnamed expressed sequence tags (ESTs) (5). Although diverse patterns of expression were observed, the orderly choreography of the expression program became apparent when the results were analyzed by a clustering and display method developed in our laboratory for analyzing genome-wide gene expression data (6). An example of such an analysis, here applied to a subset of 517 genes whose expression changed substantially in response to serum (7), is shown in Fig. 2. The entire detailed data set underlying Fig. 2 is available as a tab-delimited table (in cluster order) at the Science Web site (www.sciencemag.org/feature/data/984559.shl). In addition, the entire, larger data set for the complete set of genes analyzed in this experiment can be found at a Web site maintained by our laboratory (genome-www.stanford.edu/serum) (8).

One measure of the reliability of the changes we observed is inherent in the expression profiles of the genes. For most genes whose expression levels changed, we could see a gradual change over a few time points, which thus effectively provided independent measurements for almost all of the observations. An additional check was provided by the inclusion of duplicate and, in a few cases, multiple array elements representing the same gene for about 5% of the genes included in this microarray. In addition, three independent hybridizations to different microarrays with mRNA samples from cells harvested 8 hours after serum addition showed good correlation (Fig. 1). As an independent test, we measured the expression levels of several genes using the TaqMan 5' nuclease fluorigenic quantitative polymerase chain reaction (PCR) assay (9). The expression profiles of the genes, as measured by these two independent methods, were very similar (Fig. 3) (10).

The transcriptional response of fibroblasts to serum was extremely rapid. The immediate response to serum stimulation was dominated by genes that encode transcription factors and other proteins involved in signal transduction. The mRNAs for several genes [including c-FOS, JUN B, and mitogen-activated protein (MAP) kinase phosphatase–1 (MKP1)] were detectably induced within 15 min after serum stimulation (Fig. 4, A and B). Fifteen of the genes that were observed to be induced by serum encode known or suspected regulators of transcription (Fig. 4B). All but one were immediate-early genes—their induction was not inhibited by cycloheximide (11). This class of genes could be distinguished into those whose induction was transient (Fig. 2, cluster E) and those whose mRNA levels remained induced for much longer (Fig. 2, clusters I and J). Some features of the immediate response appeared to be directed at adaptation to the initiating signals. We observed a marked induction of mRNA encoding MKP1, a dual-specificity phosphatase that modulates the activity of the ERK1 and ERK2 MAP kinases (12). The coincidence of the peak of expression of genes in cluster E (Fig. 2) with that of MKP1 (Fig. 4A) suggests the possibility

Fig. 2. Cluster image showing the different classes of gene expression profiles. Five hundred seventeen genes whose mRNA levels changed in response to serum stimulation were selected (7). This subset of genes was clustered hierarchically into groups on the basis of the similarity of their expression profiles by the procedure of Eisen et al. (6). The expression pattern of each gene in this set is displayed here as a horizontal strip. For each gene, the ratio of mRNA levels in fibroblasts at the indicated time after serum stimulation ("unsync" denotes exponentially growing cells) to its level in the serum-deprived (time zero) fibroblasts is represented by a color, according to the color scale at the bottom. The graphs show the average expression profiles for the genes in the corresponding "cluster" (indicated by the letters A to J and color coding). In every case examined, when a gene was represented by more than one array element, the multiple representations in this set were seen to have identical or very similar expression profiles, and the profiles corresponding to these independent measurements clustered either adjacent or very close to each other, pointing to the robustness of the clustering algorithm in grouping genes with very similar patterns of expression.

that continued activity f the MAP kinase pathway is required to maintain induction of these genes but not of those with sustained expression (clusters I and J). The gene encoding a second member f the dual-specificity MAP kinase phosphatase family, known as dual-specificity protein phosphatase 6/pyst2, was induced later, at about 4 hours after serum stimulation. Genes encoding diverse other proteins with roles in signal transduction, ranging from cell-surface receptors [for example, the sphingosine 1-phosphate receptor (EDG-1), the vascular endothelial growth factor receptor, and the type II BMP receptor] to regulators of G-protein signaling (for example, NET1/p115 rho GEF) to DNA-binding transcription factors, were induced by serum (Fig. 4A).

The reprogramming of the regulatory circuits in response to serum involved not only induction of transcription factors but also reduced expression of many transcriptional regulators—some of which may play roles in maintaining the cells in G_0 or in priming them to react to wounding (Fig. 4C). Perhaps as a consequence of the historical focus on genes induced by serum stimulation of fibroblasts, the set of transcription factors whose expression diminished upon serum stimulation has been less well characterized.

Genes known or likely to be involved in controlling and mediating the proliferative response showed distinctive patterns of regulation. Several genes whose products inhibit progression of the cell-division cycle, such as p27 Kip1, p57 Kip2, and p18, were expressed in the quiescent fibroblasts and down-regulated before the onset of cell division. The nadir in the mRNA levels for these genes occurred between 6 and 12 hours after serum stimulation (Fig. 5A), coincident with the passage of the fibroblasts through G_1. The levels of the transcript encoding the WEE1-like protein kinase, which is believed to inhibit mitosis by phosphorylation of Cdc2, diminished between 4 and 8 to 12 hours after serum addition (Fig. 5A), well

before the onset of M phase at around 16 hours, raising the possibility of an additional role for Wee1 in an earlier stage of the cell cycle or in regulating the G_0 to G_1 transition. Several genes induced in the first few hours after serum stimulation, such as the helix-loop-helix proteins ID2 and ID3 and EST AA016305, a gene with homology to G_1-S cyclins, are candidates for roles in promoting the exit from G_0.

Genes involved in mediating progression through the cell cycle were characterized by a distinctive pattern of expression (Fig. 2, cluster D), reflecting the coincidence of their expression with the reentry of the stimulated fibroblasts into the cell-division cycle. The stimulated fibroblasts replicated their DNA about 16 hours after serum treatment. This timing was reflected by the induction of mRNA encoding both subunits of ribonucleotide reductase and PCNA, the processivity factor for DNA polymerase epsilon and delta. Cyclin A, Cyclin B1, Cdc2, and CDC28 kinase, regulators of passage through the S phase and the transition from G_2 to M phase, were induced at about 16 to 20 hours after serum addition. The kinase in the Cyclin B1–CDK pair needs to be activated by phosphorylation. The gene encoding Cyclin-dependent kinase 7 (CDK7: a homolog of Xenopus MO15 cdk-activating kinase) was induced in parallel with the Cdc2 and Cdc28 kinases (Fig. 5A), suggesting a potential role for CDK7 in mediating M phase. DNA topoisomerase II α, required for chromosome segregation at mitosis; Mad2, a component of the spindle checkpoint that prevents completion of mitosis (anaphase) if chromosomes are not attached to the spindle; and the kinetochore protein CENP-F all showed a similar expression profile.

In the hours after the serum stimulus, one of the most striking features of the unfolding transcriptional program was the appearance of numerous genes with known roles in processes relevant to the physiology of wound healing.

These included both genes involved in the direct role played by fibroblasts in remodeling of the clot and the extracellular matrix and, more notably, genes encoding proteins involved in intercellular signaling (Fig. 5). Genes induced in this program encode products that can (i) participate in the dynamic process of clotting, clot dissolution, and remodeling and perhaps contribute to hemostasis by promoting local vasoconstriction (for example, endothelin-1); (ii) promote chemotaxis and activation of neutrophils (for example, COX2) and recruitment and extravasation of monocytes and macrophages (for example, MCP1); (iii) promote chemotaxis and activation of T lymphocytes [for example, interleukin-8 (IL-8)] and B lymphocytes (for example, ICAM-1), thus providing both innate and antigen-specific defenses against wound infection and recruiting the phagocytic cells that will be required to clear out the debris during remodeling of the wound; (iv) promote angiogenesis and neovascularization (for example, VEGF) through newly forming tissue; (v) promote migration and proliferation of fibroblasts (for example, CTGF) and their differentiation into myofibroblasts (for example, Vimentin); and (vi) promote migration and proliferation of keratinocytes, leading to reepithelialization of the wound (for example, FGF7), and promote proliferation of melanocytes, perhaps contributing to wound hyperpigmentation (for example, FGF2).

Coordinated regulation of groups of genes whose products act at different steps in a common process was a recurring theme. For example, Furin, a prohormone-processing protease required for one of the processing steps in the generation of active endothelin, was induced in parallel with induction of the gene encoding the precursor of endothelin-1 (Fig. 5E) (13). Conversely, expression of CALLA/CD10, a membrane metalloprotease that degrades endothelin-1 and other peptide mediators of acute inflammation, was re-

Fig. 3. Independent verification of microarray quantitation. Relative mRNA levels of the indicated genes (Mast, mast/stem cell growth factor receptor) were measured with the TaqMan 5′ nuclease fluorigenic quantitative PCR assay (9) (left) in the same samples that were used to prepare probes for microarray hybridizations (right). Data from the TaqMan analysis were normalized to mRNA concentrations and plotted relative to the level at time zero, so that the results could be compared with those from the microarray hybridizations. In general, quantitation with the two methods gave very similar results (10).

duced. A second example is provided by a set of five genes involved in the biosynthesis of cholesterol (Fig. 5I). The mRNAs encoding each f these enzymes showed sharply diminished expression beginning 4 to 6 hours after serum stimulation of fibroblasts. A likely explanation for the coordinated down-regulation of the cholesterol biosynthetic pathway is that serum provides cholesterol to fibroblasts through low-density lipoproteins, whereas in the absence of the cholesterol provided by serum, endogenous cholesterol biosynthesis in fibroblasts is required.

Many of the previously studied genes that we observed to be regulated in this program have no recognized role in any aspect of wound healing or fibroblast proliferation. Their identification in this study may therefore point to previously unknown aspects of these processes. A few selected genes in this group are shown in Fig. 5H. The stanniocalcin gene, for example (Fig. 5H), encodes a secreted protein without a clearly identified function in human cells (14, 15). Its induction in serum-stimulated fibro-

blasts suggests the possibility that it may play a role in the wound-healing process, perhaps serving as a signal in mediating inflammation or angiogenesis.

One of the most important results of this exploration was the discovery of over 200 previously unknown genes whose expression was regulated in specific temporal patterns during the response of fibroblasts to serum. For example, 13 of the 40 genes in cluster D (Fig. 2) have descriptive names that reflect their putative function. Nine of these 13 genes (69%) encode proteins that play roles in cell cycle progression, particularly in DNA replication and the G_2-M transition. This enrichment for cell cycle–related genes suggests that some of the

unnamed genes in this cluster—for example, EST W79311 and EST R13146, neither of which have sequence similarity to previously characterized genes—may represent previously unknown genes involved in this part of the cell cycle. Similarly, a remarkable fraction of genes that were grouped into cluster F on the basis of their expression profiles encoded proteins involved in intercellular signaling (Fig. 2), suggesting that a similar role should be considered for the many unnamed genes in this cluster. A disproportionately large fraction of the genes whose transcription diminished upon serum stimulation were unnamed ESTs.

Our intention was to use this experiment as a model to study the control of the transition

Fig. 4. "Reprogramming" of fibroblasts. Expression profiles of genes whose function is likely to play a role in the reprogramming phase of the response are shown with the same representation as in Fig. 2. In the cases in which a gene was represented by more than one element in the microarray, all measurements are shown. The genes were grouped into categories on the basis of our knowledge of their most likely role. Some genes with pleiotropic roles were included in more than one category.

Fig. 5. The transcriptional response to serum suggests a multifaceted role for fibroblasts in the physiology of wound healing. The features of the transcriptional program of fibroblasts in response to serum stimulation that appear to be related to various aspects of the wound-healing process and fibroblast proliferation are shown with the same convention for representing changes in transcript levels as was used in Figs. 2 and 4. (A) Cell cycle and proliferation, (B) coagulation and hemostasis, (C) inflammation, (D) angiogenesis, (E) tissue remodeling, (F) cytoskeletal reorganization, (G) reepithelialization, (H) unidentified role in wound healing, and (I) cholesterol biosynthesis. The numbers in (C) and (G) refer to genes whose products serve as signals to neutrophils (C1), monocytes and macrophages (C2), T lymphocytes (C3), B lymphocytes (C4), and melanocytes (G1).

from G_0 to a proliferating state. However, one of the defining characteristics of genome-scale expression profiling experiments is that the examination f so many diverse genes opens a window on all the processes that actually occur and not merely the single process one intended to observe. Serum, the soluble fraction of clotted blood, is normally encountered by cells in vivo in the context of a wound. Indeed, the expression program that we observed in response to serum suggests that fibroblasts are programmed to interpret the abrupt exposure to serum not as a general mitogenic stimulus but as a specific physiological signal, signifying a wound. The proliferative response that we originally intended to study appeared to be part of a larger physiological response of fibroblasts to a wound. Other features of the transcriptional response to serum suggest that the fibroblast is an active participant in a conversation among the diverse cells that work together in wound repair, interpreting, amplifying, modifying, and broadcasting signals controlling inflammation, angiogenesis, and epithelial regrowth during the response to an injury.

We recognize that these in vitro results almost certainly represent a distorted and incomplete rendering of the normal physiological response of a fibroblast to a wound. Moreover, only the responses elicited directly by exposure of fibroblasts to serum were examined. The subsequent signals from other cellular participants in the normal wound-healing process would certainly provoke further evolution of the transcriptional program in fibroblasts at the site of a wound, which this experiment cannot reveal. Nevertheless, we believe that the picture that emerged strongly suggests a much larger and richer role for the fibroblast in the orchestration of this important physiological process than had previously been suspected.

References and Notes

1. J. A. Winkles, *Prog. Nucleic Acid Res. Mol. Biol.* **58**, 41 (1998).
2. A normal human diploid fibroblast cell line derived from foreskin (ATCC CRL 2091) in passage 8 was used in these experiments. The protocol followed for growth arrest and stimulation was essentially that of (16) and (17). Cells were grown to about 60% confluence in 15-cm petri dishes in Dulbecco's minimum essential medium containing glucose (1 g/liter), the antibiotics penicillin and streptomycin, and 10% (by vol) FBS (Hyclone) that had been previously heat inactivated at 56°C for 30 min. The cells were then washed three times with the same medium lacking FBS, and low-serum medium (0.1% FBS) was added to the plates. After a 48-hour incubation, the medium was replaced with fresh medium containing 10% FBS. mRNA was isolated from several plates of cells harvested before serum stimulation; this mRNA served as the serum-starved or time-zero reference sample. Cells were harvested from batches of plates at 11 subsequent intervals (15 min, 30 min, 1, 2, 4, 6, 8, 12, 16, 20, and 24 hours) after the addition of serum. mRNA was also isolated from exponentially growing fibroblasts (not subjected to serum starvation). mRNA was isolated with the FastTrack mRNA isolation kit (Invitrogen), which involves lysis of the cells on the plate. The growth medium was removed, and the cells were quickly washed with phosphate-buffered saline at room temperature. The lysis buffer was added to the plate, transferred to tubes, and frozen in liquid nitrogen. Subsequent steps were performed according to the kit manufacturer's protocols.
3. The National Center for Biotechnology Information maintains the UniGene database as a resource for partitioning human sequences contained in GenBank into clusters representing distinct transcripts or genes (18, 19). At the time this work began, this database contained about 40,000 such clusters. We selected a subset of 10,000 of these UniGene clusters for inclusion on gene expression microarrays. UniGene clusters were included only if they contained at least one clone from the I.M.A.G.E. human cDNA collection (20), so that a physical clone could easily be obtained (all I.M.A.G.E. clones are available commercially from a number of vendors). We attempted to include as complete as possible a set of the "named" human genes (about 4000) and genes that appeared to be closely related to named genes in other organisms (about an additional 2000). The remaining 4000 clones were chosen from among the "anonymous" UniGene clusters on the basis of inclusion on the human transcript map (www.ncbi.nlm.nih.gov/SCIENCE96/) and the lack of apparent homology to any other genes in the selected set. A physical clone representing each of the selected genes was obtained from Research Genetics. This "10K set" is included in a more recent "15K set" described at www.nhgri.nih.gov/DIR/LCG/15K/HTML/p15Ktop.html. Of these clones, 472 are absent from the current edition of UniGene and were presumed to be distinct genes. The remainders represent 8141 distinct clusters, or human genes, in UniGene. These clones, thus presumed to represent 8613 different genes, were used to print microarrays according to methods described previously (21, 22).
4. One microgram of mRNA was used for making fluorescently labeled cDNA probes for hybridizing to the microarrays, with the protocol described previously (23). mRNA from the large batch of serum-starved cells was used to make cDNA labeled with Cy3. The Cy3-labeled cDNA from this batch of serum-starved cells served as the common reference probe in all hybridizations. mRNA samples from cells harvested immediately before serum stimulation, at intervals after serum stimulation, and from exponentially growing cells were used to make cDNA labeled with Cy5. Ten micrograms of yeast tRNA, 10 µg of polydeoxyadenylic acid, and 20 µg of human CoT1 DNA (Gibco-BRL) were added to the mixture of labeled probes in a solution containing 3× standard saline citrate (SSC) and 0.3% SDS and allowed to prehybridize at room temperature for 30 min before the probe was added to the surface of the microarray. Hybridizations, washes, and fluorescent scans were performed as described previously (23, 24). All measurements, totaling more than 180,000 differential expression measurements, were stored in a computer database for analysis and interpretation.
5. The nominal identities of a number of cDNAs (currently about 3750) on the microarray were verified by sequencing. The clones that were sequenced included many of the genes whose expression changed substantially upon serum stimulation, as well as a large number of genes whose expression did not change substantially in the course of this experiment. About 85% of the clones on the current version of this microarray that were checked by resequencing were correctly identified. In all the figures, gene names or EST numbers are given only for those genes on the microarrays whose identities were reconfirmed by resequencing. In the cases where a human gene has more than one name in the literature, we have tried to use the name that is most evocative of its presumed role in this context. The remainder of the clones have been assigned a temporary identification number (format: SID######) and a putative identity pending sequence verification. The correct identities of these genes will be posted at our Web site (genome-www.stanford.edu/serum) as they are confirmed by resequencing.
6. M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, *Proc. Natl. Acad. Sci. U.S.A.* **95**, 14863 (1998).
7. Genes were selected for this analysis if either (i) their expression level deviated from that in quiescent fibroblasts by at least a factor of 2.20 in at least two of the samples from serum-stimulated cells or (ii) the standard deviation for the set of 13 values of $\log_2$(expression ratio) measured for the gene in this time course exceeded 0.7. In addition, observations in which the pixel-by-pixel correlation coefficients for the Cy3 and Cy5 fluorescence signals measured in a given array element were less than 0.6 were excluded. This selection criteria yielded a computationally manageable number of genes while minimizing the number of genes that were included because of noise in the data.
8. A more complete analysis and interpretation of the results of this experiment, as well as a searchable database, can be found at genome-www.stanford.edu/serum
9. K. J. Livak, S. J. Flood, J. Marmaro, W. Giusti, K. Deetz, *PCR Methods Appl.* **4**, 357 (1995).
10. The apparent dip in the profile of COX2 at the 2-hour time point in the microarray data appears to result from a localized area of low intensity on the corresponding array scan resulting in an underestimation of the expression ratio. The expression ratios measured for mast/stem cell growth factor receptor are somewhat lower in the microarray data. This discrepancy is probably a consequence of the conservative background subtraction method used for quantitating the signal intensities on the array scans (23). The sequences of the PCR primer pairs (5' to 3') that were used are as follows: COX2, CCGTGGCTCTCTT-GGCAG and CTAAGTTCTTTAGCACTCCTTGGCA; IL-8, CGATGCTGTGGACGTGTATC and CCATGGTTTC-ACCAAAGATG; mast/stem cell factor receptor, ACA-GAAGCCCGTGGTAGACC and GAGGCTGGGAGGAG-GAAG; B4-2, AAACCCCCCTCAGGAAAGAG and CC-ATGAACAAGCTGGCCAT; and actin, AGTACTCCGTGT-GGATCGGC and GCTGATCCACATCTGCTGGA.
11. V. R. Iyer *et al.*, unpublished data. The gene expression data for the early time points in the presence of cycloheximide will be available at our Web site (genome-www.stanford.edu/serum)
12. T. Hunter, *Cell* **80**, 225 (1995).
13. J. Leppaluoto and H. Ruskoaho, *Ann. Med.* **24**, 153 (1992).
14. A. C. Chang *et al.*, *Mol. Cell. Endocrinol.* **112**, 241 (1995).
15. K. L. Madsen *et al.*, *Am. J. Physiol.* **274**, C96 (1998).
16. W. Krek and J. A. DeCaprio, *Methods Enzymol.* **254**, 114 (1995).
17. R. A. Tobey, J. G. Valdez, H. A. Crissman, *Exp. Cell Res.* **179**, 400 (1988).
18. M. S. Boguski and G. D. Schuler, *Nature Genet.* **10**, 369 (1995).
19. G. D. Schuler, *J. Mol. Med.* **75**, 694 (1997).
20. G. Lennon, C. Auffray, M. Polymeropoulos, M. B. Soares, *Genomics* **33**, 151 (1996).
21. I.M.A.G.E. clones were amplified by PCR in 96-well format with amino-linked primers at the 5' end. Purified PCR products were suspended at a concentration of ~0.5 mg/ml in 3× SSC, and ~5 ng of each product was arrayed onto coated glass by means of procedures similar to those described previously (22). A total of 9996 elements were arrayed onto an area of 1.8 cm by 1.8 cm with the elements spaced 175 µm apart. The microarrays were then postprocessed to fix the DNA to the glass surface before hybridization with a procedure similar to previously described methods (22).
22. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
23. J. L. DeRisi, V. R. Iyer, P. O. Brown, *ibid.* **278**, 680 (1997).
24. J. DeRisi *et al.*, *Nature Genet.* **14**, 457 (1996).
25. We thank E. Chung for help with sequencing, A. Alizadeh for help with sequence verification, K. Ranade for advice on the TaqMan assay, and J. DeRisi and other members of the P.O.B. and D.B. labs for discussions. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450) and the National Cancer Institute (NIH CA 77097). V.R.I. was supported in part by an Institutional Training Grant in Genome Sciences (T32 HG00044) from the NHGRI. M.B.E. is an Alfred E. Sloan Foundation Postdoctoral Fellow in Computational Molecular Biology, and D.T.R is a Walter and Idun Berry Fellow. P.O.B. is an Associate Investigator of the Howard Hughes Medical Institute.

13 August 1998; accepted 13 November 1998

Systematic variation in gene expression patterns in human cancer cell lines

Douglas T. Ross[1], Uwe Scherf[5], Michael B. Eisen[2], Charles M. Perou[2], Christian Rees[2], Paul Spellman[2], Vishwanath Iyer[1], Stefanie S. Jeffrey[3], Matt Van de Rijn[4], Mark Waltham[5], Alexander Pergamenschikov[2], Jeffrey C.F. Lee[6], Deval Lashkari[7], Dari Shalon[6], Timothy G. Myers[8], John N. Weinstein[5], David Botstein[2] & Patrick O. Brown[1,9]

We used cDNA microarrays to explore the variation in expression of approximately 8,000 unique genes among the 60 c ll lines used in the National Cancer Institute's screen for anti-cancer drugs. Classification of the cell lines based solely on the observed patterns of gene expression revealed a correspondence to the ostensible origins of the tumours from which the cell lines were derived. The consistent relationship between the gene expression patterns and the tissue of origin allowed us to recognize outliers whose previous classification appeared incorrect. Specific features of the gene expression patterns appeared to be related to physiological properties of the cell lines, such as their doubling time in culture, drug metabolism or the interferon response. Comparison of gene expression pat- t rns in the cell lines to those observed in normal breast tissue or in breast tumour specimens revealed features of the expression patterns in the tumours that had recognizable counterparts in specific cell lines, reflecting the tumour, stromal and inflammatory components of the tumour tissue. These results provided a novel molecular characterization of this important group of human cell lines and their relationships to tumours *in vivo*.

Intr duction

Cell lines derived from human tumours have been extensively used as experimental models of neoplastic disease. Although such cell lines differ from both normal and cancerous tissue, the inaccessibility of human tumours and normal tissue makes it likely that such cell lines will continue to be used as experimental models for the foreseeable future. The National Cancer Institute's Developmental Therapeutics Program (DTP) has carried out intensive studies of 60 cancer cell lines (the NCI60) derived from tumours from a variety of tissues and organs[1–4]. The DTP has assessed many molecular features of the cells related to cancer and chemotherapeutic sensitivity, and has measured the sensitivities of these 60 cell lines to more than 70,000 different chemical compounds, including all common chemotherapeutics (http://dtp.nci.nih.gov). A previous analysis of these data revealed a connection between the pattern of activity of a drug and its method of action. In particular, there was a tendency for groups of drugs with similar patterns of activity to have related methods of action[3,5–7].

We used DNA microarrays to survey the variation in abundance of approximately 8,000 distinct human transcripts in these 60 cell lines. Because of the logical connection between the function of a gene and its pattern of expression, the correlation of gene expression patterns with the variation in the phenotype of the cell can begin the process by which the function of a gene can be inferred. Similarly, the patterns of expression of known genes can

reveal novel phenotypic aspects of the cells and tissues studied[8–10]. Here we present an analysis of the observed patterns of gene expression and their relationship to phenotypic properties of the 60 cell lines. The accompanying report[11] explores the relationship between the gene expression patterns and the drug sensitivity profiles measured by the DTP. The assessment of gene expression patterns in a multitude of cell and tissue types, such as the diverse set of cell lines we studied here, under diverse conditions *in vitro* and *in vivo*, should lead to increasingly detailed maps of the human gene expression program and provide clues as to the physiological roles of uncharacterized genes[11–16]. The databases, plus tools for analysis and visualization of the data, are available (http://genome-www.stanford.edu/nci60 and http://discover.nci.nih.gov).

Results

We studied gene expression in the 60 cell lines using DNA microarrays prepared by robotically spotting 9,703 human cDNAs on glass microscope slides[17,18]. The cDNAs included approximately 8,000 different genes: approximately 3,700 represented previously characterized human proteins, an additional 1,900 had homologues in other organisms and the remaining 2,400 were identified only by ESTs. Due to ambiguity of the identity of the cDNA clones used in these studies, we estimated that approximately 80% of the genes in these experiments were correctly identified. The identities of approximately 3,000 cDNAs

Departments of [1]Biochemistry, [2]Genetics, [3]Surgery and [4]Pathology, Stanford University School of Medicine, Stanford, California, USA. [5]Laboratory of Molecular Pharmacology, Division of Basic Sciences, National Cancer Institute, National Institutes of Health, Bethesda, Maryland, USA. [6]Incyte Pharmaceuticals, Fremont, California, USA. [7]Genometrix Inc., The Woodlands, Texas, USA. [8]Information Technology Branch, Developmental Therapeutics Program, Division of Cancer Treatment and Diagnosis, National Cancer Institute, National Institutes of Health, Rockville, Maryland, USA. [9]Howard Hughes Medical Institute, Stanford University School of Medicine, Stanford, California, USA. Correspondence should be addressed to P.O.B. (e-mail: pbrown@cmgm.stanford.edu) or J.N.W. (e-mail: Weinstein@dtpax2.ncifcrf.gov).
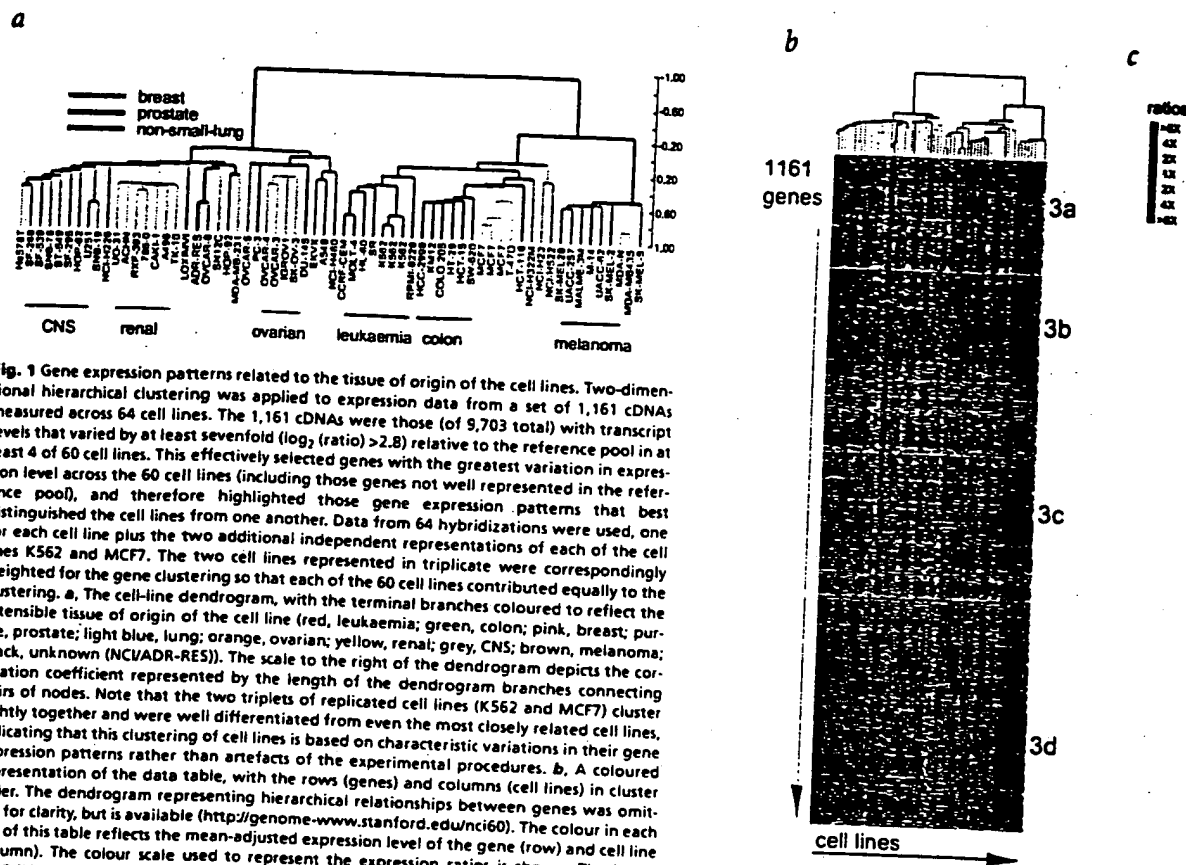
a

b

c

Fig. 1 Gene expression patterns related to the tissue of origin of the cell lines. Two-dimensional hierarchical clustering was applied to expression data from a set of 1,161 cDNAs measured across 64 cell lines. The 1,161 cDNAs were those (of 9,703 total) with transcript levels that varied by at least sevenfold ($\log_2$ (ratio) >2.8) relative to the reference pool in at least 4 of 60 cell lines. This effectively selected genes with the greatest variation in expression level across the 60 cell lines (including those genes not well represented in the reference pool), and therefore highlighted those gene expression patterns that best distinguished the cell lines from one another. Data from 64 hybridizations were used, one for each cell line plus the two additional independent representations of each of the cell lines K562 and MCF7. The two cell lines represented in triplicate were correspondingly weighted for the gene clustering so that each of the 60 cell lines contributed equally to the clustering. *a*, The cell-line dendrogram, with the terminal branches coloured to reflect the ostensible tissue of origin of the cell line (red, leukaemia; green, colon; pink, breast; purple, prostate; light blue, lung; orange, ovarian; yellow, renal; grey, CNS; brown, melanoma; black, unknown (NCI/ADR-RES)). The scale to the right of the dendrogram depicts the correlation coefficient represented by the length of the dendrogram branches connecting pairs of nodes. Note that the two triplets of replicated cell lines (K562 and MCF7) cluster tightly together and were well differentiated from even the most closely related cell lines, indicating that this clustering of cell lines is based on characteristic variations in their gene expression patterns rather than artefacts of the experimental procedures. *b*, A coloured representation of the data table, with the rows (genes) and columns (cell lines) in cluster order. The dendrogram representing hierarchical relationships between genes was omitted for clarity, but is available (http://genome-www.stanford.edu/nci60). The colour in each cell of this table reflects the mean-adjusted expression level of the gene (row) and cell line (column). The colour scale used to represent the expression ratios is shown. The labels '3a–3d' in (*b*) refer to the clusters of genes shown in detail in Fig. 3.

from these experiments have been sequence-verified, including all of those referred to here by name.

Each hybridization compared Cy5-labelled cDNA reverse transcribed from mRNA isolated from one of the cell lines with Cy3-labelled cDNA reverse transcribed from a reference mRNA sample. This reference sample, used in all hybridizations, was prepared by combining an equal mixture of mRNA from 12 of the cell lines (chosen to maximize diversity in gene expression as determined primarily from two-dimensional gel studies[2]). By comparing cDNA from each cell line with a common reference, variation in gene expression across the 60 cell lines could be inferred from the observed variation in the normalized Cy5/Cy3 ratios across the hybridizations.

To assess the contribution of artefactual sources of variation in the experimentally measured expression patterns, K562 and MCF7 cell lines were each grown in three independent cultures, and the entire process was carried out independently on mRNA extracted from each culture. The variance in the triplicate fluorescence ratio measurements approached a minimum when the fluorescence signal was greater than approximately 0.4% of the measurable total signal dynamic range above background in either channel of the hybridization. We selected the subset of spots for which significant signal was present in both the numerator and denominator of the ratios by this criterion to identify the best-measured spots. The pair-wise correlation coefficients for the triplicates of the set of genes that passed this quality control level (6,992 spots included for the MCF7 samples and 6,161 spots for K562) ranged from 0.83 to 0.92 (for graphs and details, see http://genome-www.stanford.edu/nci60).

To make the orderly features in the data more apparent, we used a hierarchical clustering algorithm[19,20] and a pseudo-colour visu-

alization matrix[3,21]. The object of the clustering was to group cell lines with similar repertoires of expressed genes and to group genes whose expression level varied among the 60 cell lines in a similar manner. Clustering was performed twice using different subsets of genes to assess the robustness of the analysis. In one case (Fig. 1), we concentrated on those genes that showed the most variation in expression among the 60 cell lines (1,167 total). A second analysis (Fig. 2) included all spots that were thought to be well measured in the reference set (6,831 spots).

Gene expression patterns related to the histologic origins of the cell lines

The most notable property of the clustered data was that cell lines with common presumptive tissues of origin grouped together (Figs 1*a* and 2). Cell lines derived from leukaemia, melanoma, central nervous system, colon, renal and ovarian tissue were clustered into independent terminal branches specific to their respective organ types with few exceptions. Cell lines derived from non-small lung carcinoma and breast tumours were distributed in multiple different terminal branches suggesting that their gene expression patterns were more heterogeneous.

Many of these coherent cell line clusters were distinguished by the specific expression of characteristic groups of genes (Fig. 3*a–d*). For example, a cluster of approximately 90 genes was highly expressed in the melanoma-derived lines (Fig. 3*c*). This set was enriched for genes with known roles in melanocyte biology, including tyrosinase and dopachrome tautomerase (TYR and DCT; two subunits of an enzyme complex involved in melanin synthesis[22]), MART1 (MLANA; which is being investigated as a target for immunotherapy of melanoma[23]) and S100-β (S100B; which has been used as an antigenic marker in the diagnosis of

228

article

Fig. 2 Gene expression patterns related to other cell-line phenotypes. a, We applied two-dimensional hierarchical clustering to expression data from a set of 6,831 cDNAs measured across the 64 cell lines. The 6,831 cDNAs were those with a minimum fluorescence signal intensity of approximately 0.4% of the dynamic range above background in the reference channel in each of the six hybridizations used to establish reproducibility. This effectively selected those spots that provided the most reliable ratio measurements and therefore identified a subset of genes useful for exploring patterns comprised of those whose variation in expression across the 60 cell lines was of moderate magnitude. **b,** Cluster-ordered data table. **c,** Doubling time of cell lines. Cell lines are given in cluster order. Values are plotted relative to the mean. Doubling times greater than the mean are shown in green, those with doubling time less than the mean are shown in red. **d,** Three related gene clusters that were enriched for genes whose expression level variation was correlated with cell line proliferation rate. Each of the three gene clusters (clustered solely on the basis of their expression patterns) showed enrichment for sets of genes involved in distinct functional categories (for example, ribosomal genes versus genes involved in pre-RNA splicing). **e,** Gene cluster in which all characterized and sequence-verified cDNAs encode genes known to be regulated by interferons. **f,** Gene cluster enriched for genes that have been implicated in drug metabolism (indicated by asterisks). A further property of the gene clustering evident here and in Fig. 2 is the strong tendency for redundant representations of the same gene to cluster immediately adjacent to one another, even within larger groups of genes with very similar expression patterns. In addition to illustrating the reproducibility and consistency of the measurements, and providing independent confirmation of many of our measurements, this property also demonstrates that these, and probably all, genes have nearly unique patterns of variation across the 60 cell lines. If this were not the case, and multiple genes had identical patterns of variation, we would not expect to be able to distinguish, by clustering on the basis of expression variation, duplicate copies of individual genes from the other genes with identical expression patterns.

a

breast
prostate
non-small-lung

CNS   renal   ovarian   leukaemia   colon   melanoma

b

6831 genes

c   doubling time   ratios

hours

mean →

d   proliferation cluster

e   interferon cluster

f   drug metabolism cluster

cell lines

melanoma). LOXIMVI, the seventh line designated as melanoma in the NCI60, did not show this characteristic pattern. Although isolated from a patient with melanoma, LOXIMVI has previously been noted to lack melanin and other markers useful for identification of melanoma cells[1].

Paradoxically, two related cell lines (MDA-MB435 and MDA-N), which were derived from a single patient with breast cancer and have been conventionally regarded as breast cancer cell lines, shared expression of the genes associated with melanoma. MDA-MB435 was isolated from a pleural effusion in a patient with metastatic ductal adenocarcinoma of the breast[24,25]. It remains possible that the origin of the cell line was a breast cancer, and that its gene expression pattern is related to the neuroendocrine features of some breast cancers[26]. But our results suggest that this cell line may have originated from a melanoma, raising the possibility that the patient had a co-existing occult melanoma.

The higher-level organization of the cell-line tree—in which groups span cell lines from different tissue types—also reflected shared biological properties of the tissues from which the cell lines were derived. The carcinoma-derived cell lines were divided into major branches that separated those that expressed genes characteristic of epithelial cells from those that expressed genes more typical of stromal cells. A cluster of genes is shown (Fig. 3b) that is most strongly expressed in cell lines derived from colon carcinomas, six of seven ovarian-derived cell lines and the two breast cancer lines positive for the oestrogen receptor. The named genes in this cluster have been implicated in several aspects of epithelial cell biology[27]. The cluster was enriched for genes whose products are known to localize to the basolateral membrane of epithelial cells, including those encoding components of adherens complexes (for example, desmoplakin (DSP), periplakin (PPL) and plakoglobin (JUP)), an epithelial-expressed cell-cell adhesion molecule (M4S1) and a sodium/hydrogen ion exchanger[28–31] (SLC9A1). It also contained genes that encode putative transcriptional regulators of epithelial morphogenesis, a human homologue of a *Drosophila melanogaster* epithelial-expressed tumour suppressor (LLGL1) and a homeobox gene thought to control calcium-mediated adherence in epithelial cells[32,33] (MSX2).

In contrast, a separate, major branch of the cell-line dendrogram (Fig. 1a) included all glioblastoma-derived cell lines, all renal-cell-carcinoma–derived cell lines and the remaining carcinoma-derived lines. The characteristic set of genes expressed in this cluster included many whose products are involved in stromal cell functions (Fig. 3d). Indeed, the two cell lines originally described as 'sarcoma-like' in appearance (Hs578T, breast carcinosarcoma, and SF539, gliosarcoma) expressed most of these genes[34,35]. Although no single gene was uniformly characteristic of this cluster, each cell line showed a distinctive pattern of expression of genes encoding proteins with roles in synthesis or modification of the extracellular matrix (for example, caldesmon (CALD1), cathepsins, thrombospondin (THBS), lysyl oxidase (LOX) and collagen subtypes). Although the ovarian and most non-small-cell-lung–derived carcinomas expressed genes characteristic of both epithelial cells and stromal cells, they probably clustered with the CNS and renal cell carcinomas in this analysis because genes characteristically expressed in stromal cells were more abundantly represented in this gene set.

Physiological variation reflected in gene expression patterns

A cluster diagram of 6,831 genes (Fig. 2) is useful for exploring clusters of genes whose variation in mRNA levels was not obviously attributable to cell or tissue type. We identified some gene clusters that were enriched for genes involved in specific cellular

processes; the variation in their expression levels may reflect corresponding differences in activity of these processes in the cell lines. For example, a cluster of 1,159 genes (Fig. 2a) included many whose products are necessary for progression through the cell cycle (such as CCNA1, MCM106 and MAD2L1), RNA processing and translation machinery (such as RNA helicases, hnRNPs and translation elongation factors) and traditional pathologic markers used to identify proliferating cells (MKI67). Within this large cluster were smaller clusters enriched for genes with more specialized roles. One cluster was highly enriched for numerous ribosomal genes, whereas another was more enriched for genes encoding RNA-splicing factors. The variation in expression of these ribosomal genes was significantly correlated with variation in the cell doubling time (correlation coefficient of 0.54), supporting the notion that the genes in this cluster were regulated in relation to cell proliferation rate or growth rate in these cell lines.

In a smaller gene cluster (Fig. 2d), all of the named genes were previously known to be regulated by interferons[13,36]. Additional groups of interferon-regulated genes showed distinct patterns of expression (data not shown), suggesting that the NCI60 cell lines exhibited variation in activity of interferon-response pathways, which was reflected in gene expression patterns[36].

Another cluster (Fig. 2e) contained several genes encoding proteins with possible interrelated roles in drug metabolism, including glutamate-cysteine ligase (GLCLC, the enzyme responsible for the rate limiting step of glutathione synthesis), thioredoxin (TXN) and thioredoxin reductase (TXNRD1; enzymes involved in regulating redox state in cells), and MRP1 (a drug transporter known to efficiently transport glutathione-conjugated compounds[37]). The elevated expression of this set of genes in a subset of these cell lines may reflect selection for resistance to chemotherapeutics.

Cell lines facilitate interpretation of gene expression patterns in complex clinical samples

Like many other types of cancer, tumours of the breast typically have a complex histological organization, with connective tissue and leukocytic infiltrates interwoven with tumour cells. To explore the possibility that variation in gene expression in the tumour cell lines might provide a framework for interpreting the expression patterns in tumour specimens, we compared RNA isolated from two breast cancer biopsy samples, a sample of normal breast tissue and the NCI60 cell lines derived from breast cancers (excluding MDA-MB-435 and MDA-N) and leukaemias (Fig. 4). This clustering highlighted features of the gene expression pattern shared between the cancer specimens and individual cell lines derived from breast cancers and leukaemias.

The genes encoding keratin 8 (KRT8) and keratin 19 (KRT19), as well as most of the other 'epithelial' genes defined in the complete NCI60 cell line cluster, were expressed in both of the biopsy samples and the two breast-derived cell lines, MCF-7 and T47D, expressing the oestrogen receptor, suggesting that these transcripts originated in tumour cells with features similar to those of luminal epithelial cells (Fig. 5a). Expression of a set of genes characteristic of stromal cells, including collagen genes (COL3A1, COL5A1 and COL6A1) and smooth muscle cell markers (TAGLN), was a feature shared by the tumour sample and the stromal-like cell lines Hs578T and BT549 (Fig. 5b). This feature of the expression pattern seen in the tumour samples is likely to be due to the stromal component of the tumour. The tumours also shared expression of a set of genes (Fig. 5c) with the multiple myeloma cell line (RPMI-8226), notably including immunoglobulin genes, consistent with the presence of B cells in the tumour (this was confirmed by staining with anti-

a leukaemia cluster (6 ESTs)

b epithelial cluster (15 ESTs)

c melanoma cluster (16 ESTs)

d mesenchymal cluster (67 ESTs)

Fig. 3 Gene clusters related to tissue characteristics in the cell lines. Enlargements of the regions of the cluster diagram in Fig. 1 showing gene clusters enriched for genes expressed in cell lines of ostensibly similar origins. *a*, Cluster of genes highly expressed in the leukaemia-derived cell lines. Two sub-clusters distinguish genes that were expressed in most leukaemia-derived lines from those expressed exclusively in the eryroblastoid line, K562 (note that the triplicate clones cluster together). *b*, Cluster of genes highly expressed in all colon (7/7) cell lines and all breast-derived cell lines positive for the oestrogen receptor (2/2). This set of genes was also moderately expressed in most ovarian lines (5/6) and some non-small-cell-lung (4/6) lines, but was expressed at a lower level in all renal-cancer–derived lines. *c*, Cluster of genes highly expressed in most melanoma-derived lines (6/7) and two related lines ostensibly derived from breast cancer (MDA-MB435 and MDA-N). *d*, Cluster of genes highly expressed in all glioblastoma (6/6) lines and most lines derived from renal-cell carcinoma (7/8), and more moderately expressed in a subset of carcinoma-derived lines. In all panels, names are shown only for all known genes whose identities were independently re-verified by sequencing. The number of sequence-validated ESTs within the cluster is indicated below the cluster in parentheses. The position of gene names in the adjacent list only approximates their position in the cluster diagram as indicated by the lines connecting the colour chart with the gene list. Complete cluster images with all gene names and accession numbers are available (http://genome-www.stanford.edu/nci60).

a

1824 genes

leukocyte cluster

b

stromal cluster

epithelial cluster

ratios
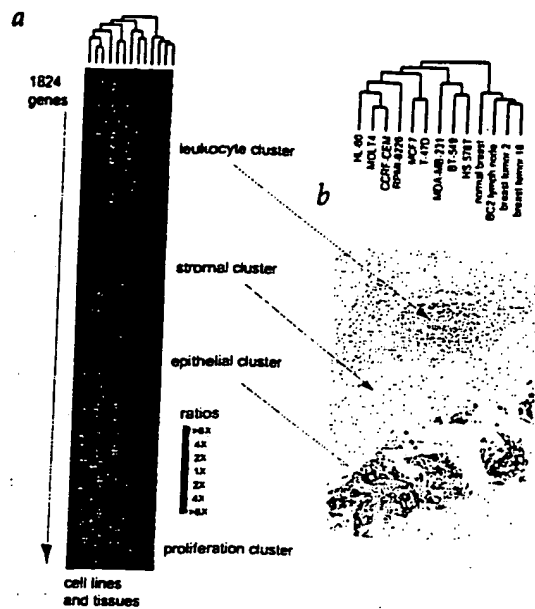
proliferation cluster

cell lines and tissues

Fig. 4 Comparison of the gene expression patterns in clinical breast cancer specimens and cultured breast cancer and leukaemia cell lines. *a*, Two-dimensional hierarchical clustering applied to gene expression data for two breast cancer specimens, a lymph node metastasis from one patient, normal breast and the NCI60 breast and leukaemia-derived cell lines. The gene expression data from tissue specimens was clustered along with expression data from a subset of the NCI60 cell lines to explore whether features of expression patterns observed in specific lines could be identified in the tissue samples. Labels indicate gene clusters (shown in detail in Fig. 5) that may be related to specific cellular components of the tumour specimens. *b*, Breast cancer specimen 16 stained with anti-keratin antibodies, showing the complex mix of cell types characteristically found in breast tumours. The arrows highlight the different cellular components of this tissue specimen that were distinguished by the gene expression cluster analysis (Fig. 5).

immunoglobulin antibodies; data not shown). Therefore, distinct sets of genes with co-varying expression among the samples (Fig. 4, arrow) appear to represent distinct cell types that can be distinguished in breast cancer tissue. A fourth cluster of genes, more highly expressed in all of the cell lines than in any of the clinical specimens, was enriched for genes present in the 'proliferation' cluster described above (Fig. 5*d*). The variation in expression of these genes likely paralleled the difference in proliferation rate between the rapidly cycling cultured cell lines and the much more slowly dividing cells in tissues.

Discussion

Newly available genomics tools allowed us to explore variation in gene expression on a genomic scale in 60 cell lines derived from diverse tumour tissues. We used a simple cluster analysis to identify the prominent features in the gene expression patterns that appeared to reflect 'molecular signatures' of the tissue from which the cells originated. The histological characteristics of the cell lines that dominated the clustering were pervasive enough that similar relationships were revealed when alternative subsets of genes were selected for analysis. Additional features of the expression pattern may be related to variation in physiological attributes such as proliferation rate and activity of interferon-response pathways.

The properties of the tumour-derived cell lines in this study have presumably all been shaped by selection for resistance to host defences and chemotherapeutics and for rapid proliferation in the tissue culture environment of synthetic growth media, fetal bovine serum and a polystyrene substratum. But the primary identifiable factor accounting for variation in gene expression patterns among these 60 cell lines was the identity of the tissue from which each cell line was ostensibly derived. For most of the cell lines we examined, neither physiological nor experimental adaptation for growth in culture was sufficient to overwrite the gene expression programs established during differentiation *in vivo*. Nevertheless, the prominence of mesenchymal features in the cell lines isolated from glioblastomas and carcinomas may reflect a selection for the relative ease of establishment of cell lines expressing stromal characteristics, perhaps combined with physiological adaptation to tissue culture conditions[38–40].

Biological themes linking genes with related expression patterns may be inferred in many cases from the shared attributes of known genes within the clusters. Uncharacterized cDNAs are likely to encode proteins that have roles similar to those of the known gene products with which they appear to be co-regulated. Still, for several clusters of genes, we were unable to discern a common theme linking the identified members of the cluster. Further exploration of their variation in expression under more diverse conditions and more comprehensive investigation of the physiology of the NCI60 cells may provide insight[10]. The relationship of the gene expression patterns to the drug sensitivity patterns measured by the DTP is an example of linking variation in gene expression with more subtle and diverse phenotypic variation[11].

The patterns of gene expression measured in the NCI60 cell lines provide a framework that helps to distinguish the cells that express specific sets of genes in the histologically complex breast cancer specimens[41]. Although it is now feasible to analyse gene expression in micro-dissected tumour specimens[42,43], this observation suggests that it will be possible to explore and interpret some of the biology of clinical tumour samples by sampling them intact. As is useful in conventional morphological pathology, one might be able to observe interactions between a tumour and its microenvironment in this way. These relationships will be clarified by suitable analysis of gene expression patterns from intact as well as dissected tumours[12,14,15,41].

Methods

(100 µl) was purified by gel exclusion, concentrated and resuspended in 3×SSC (10 µl). The PCR products were then printed on treated glass microscope slides using a robot with four printing tips. Detailed protocols for assembling and operating a microarray printer, and printing and experimental application of DNA microarrays are available (http://cmgm. stanford.edu/pbrown).

Preparation of mRNA and reference pool. Cell lines were grown from NCI DTP frozen stocks in RPMI-1640 supplemented with phenol red, glutamine (2 mM) and 5% fetal calf serum. To minimize the contribution of variations in culture conditions or cell density to differential gene expression, we grew each cell line to 80% confluence and isolated mRNA 24 h after transfer to fresh medium. The time between removal from the incubator and lysis of the cells in RNA stabilization buffer was minimized (<1 min). Cells were lysed in buffer containing guanidium isothiocyanate and total RNA was purified with the RNeasy purification kit (Qiagen). We purified mRNA as needed

using a poly(A) purification kit (Oligotex, Qiagen) according to the manufacturer's instructions. Denaturing agarose gel electrophoresis assessed the integrity and relative contamination of mRNA with ribosomal RNA.

The breast tumours were surgically excised from patients and rapidly transported to the pathology laboratory, where samples for microarray analysis were quickly frozen in liquid nitrogen and stored at –80 °C until use. A frozen tumour specimen was removed from the freezer, cut into small pieces (~50–100 mg each), immediately placed into 10–12 ml of Trizol reagent (Gibco-BRL) and homogenized using a PowerGen 125 Tissue Homogenizer (Fisher Scientific), starting at 5,000 r.p.m. and gradually increasing to ~20,000 r.p.m. over a period of 30–60 s. We processed the Trizol/tumour homogenate as described in the Trizol protocol, including an initial step to remove fat. Once total RNA was obtained, we isolated mRNA with a FastTrack 2.0 kit (Invitrogen) using the manufacturer's protocol for isolating mRNA starting from total RNA. The normal breast samples were obtained from Clontech.

Fig. 5 Histologic features of breast cancer biopsies can be recognized and parsed based on gene expression patterns. Enlargements of the regions of the cluster diagram in Fig. 4 showing gene clusters enriched for genes expressed in different cell types in the breast cancer specimens, as distinguished by clustering with the cultured cell lines. *a*, A cluster including many genes characteristic of epithelial cells expressed in cell lines (T47D and MCF7) derived from breast cancer positive for the oestrogen receptor and tumours. *b*, Genes expressed in cell lines derived from breast cancer with stromal cell characteristics (Hs578T and BT549) and tumour specimens. Expression of these genes in the tumour samples may reflect the presence of myofibroblasts in the cancer specimen stroma. *c*, Genes expressed in leukocyte-derived cell lines, showing common leukocyte, and separate 'myeloid' and 'B-cell', gene clusters. *d*, Genes that were relatively highly expressed in all cell lines compared with the tumour specimens and normal breast. The higher expression of this set of genes involved in cell cycle transit in the cell lines is likely to reflect the higher proliferative rate of cells cultured in the presence of serum compared with the average proliferation rate of cells in the biopsied tissue.

We combined mRNA from the following cells in equal quantities to make the reference pool: HL-60 (acute myeloid leukaemia) and K562 (chronic myeloid leukaemia); NCI-H226 (non-small-cell-lung); COLO 205 (colon); SNB-19 (central nervous system); LOX-IMVI (melanoma); OVCAR-3 and OVCAR-4 (ovarian); CAKI-1 (renal); PC-3 (prostate); and MCF7 and Hs578T (breast). The criterion for selection of the cell lines in the reference are described in detail in the accompanying manuscript[12].

Doubling-time calculations. We calculated doubling times based on routine NCI60 cell line compound screening data; and they reflect the doubling times for cells inoculated into 96-well plates at the screening inoculation densities and grown in RPMI 1640 medium supplemented with 5% fetal bovine serum for 48 h. We measured cell populations using sulforhodamine B optical density measurement assay. The doubling time constant k was calculated using the equation: $N/No = e^{kt}$, where No is optical density for control (untreated) cells at time zero, N is optical density for control cells after 48-h incubation, and t is 48 h. The same equation was then used with the derived k to calculate the doubling time t by setting $N/No = 2$. For a given cell line, we obtained No and N values by averaging optical densities (N>6,000) obtained for each cell line for a year's screening. Data and experimental details are available (http://dtp.nci.nih.gov).

Preparation and hybridization of fluorescent labelled cDNA. For each comparative array hybridization, labelled cDNA was synthesized by reverse transcription from test cell mRNA in the presence of Cy5-dUTP, and from the reference mRNA with Cy3-dUTP, using the Superscript II reverse-transcription kit (Gibco-BRL). For each reverse transcription reaction, mRNA (2 µg) was mixed with an anchored oligo-dT (d-20T-d(AGC)) primer (4 µg) in a total volume of 15 µl, heated to 70 °C for 10 min and cooled on ice. To this sample, we added an unlabelled nucleotide pool (0.6 µl; 25 mM each dATP, dCTP, dGTP, and 15 mM dTTP), either Cy3 or Cy5 conjugated dUTP (3 µl; 1 mM; Amersham), 5×first-strand buffer (6 µl; 250 mM Tris-HCL, pH 8.3, 375 mM KCl, 15 mM MgCl$_2$), 0.1 M DTT (3 µl) and 2 µl of Superscript II reverse transcriptase (200 µl/µl). After a 2-h incubation at 42 °C, the RNA was degraded by adding 1 N NaOH (1.5 µl) and incubating at 70 °C for 10 min. The mixture was neutralized by adding of 1 N HCL (1.5 µl), and the volume brought to 500 µl with TE (10 mM Tris, 1 mM EDTA). We added Cot1 human DNA (20 µg; Gibco-BRL), and purified the probe by centrifugation in a Centricon-30 micro-concentrator (Amicon). The two separate probes were combined, brought to a volume of 500 µl, and concentrated again to a volume of less than 7 µl. We added 10 µg/µl poly(A) RNA (1 µl; Sigma) and tRNA (10 µg/µl; Gibco-BRL) were added, and adjusted the volume to 9.5 µl with distilled water. For final probe preparation, 20×SSC (2.1 µl; 1.5 M NaCl, 150 mM NaCitrate, pH 8.0) and 10% SDS (0.35 µl) were added to a total final volume of 12 µl. The probes were denatured by heating for 2 min at 100 °C, incubated at 37 °C for 20–30 min, and placed on the array under a 22 mm×22 mm glass coverslip. We incubated slides overnight at 65 °C for 14–18 h in a custom slide chamber with humidity maintained by a small reservoir of 3×SSC. Arrays were washed by submersion and agitation for 2–5 min in 2×SSC with 0.1% SDS, followed by 1×SSC and then 0.1×SSC. The arrays were "spun dry" by centrifugation for 2 min in a slide-rack in a Beckman GS-6 tabletop centrifuge in Microplus carriers at 650 r.p.m. for 2 min.

Array quantitation and data processing. Following hybridization, arrays were scanned using a laser-scanning microscope (ref. 17; http://cmgm.stanford.edu/pbrown). Separate images were acquired for Cy3 and Cy5. We carried out data reduction with the program ScanAlyze (M.B.E., available

at http://rana.stanford.edu/software). Each spot was defined by manual positioning of a grid of circles over the array image. For each fluorescent image, the average pixel intensity within each circle was determined, and a local background was computed for each spot equal to the median pixel intensity in a square of 40 pixels in width and height centred on the spot centre, excluding all pixels within any defined spots. Net signal was determined by subtraction of this local background from the average intensity for each spot. Spots deemed unsuitable for accurate quantitation because of array artefacts were manually flagged and excluded from further analysis. Data files generated by ScanAlyze were entered into a custom database that maintains web-accessible files. Signal intensities between the two fluorescent images were normalized by applying a uniform scale factor to all intensities measured for the Cy5 channel. The normalization factor was chosen so that the mean log(Cy3/Cy5) for a subset of spots that achieved a minimum quality parameter (approximately 6,000 spots) was 0. This effectively defined the signal-intensity-weighted 'average' spot on each array to have a Cy3/Cy5 ratio of 1.0.

Cluster analysis. We extracted tables (rows of genes, columns of individual microarray hybridizations) of normalized fluorescence ratios from the database. Various selection criteria, discussed in relation to each data set, were applied to select subsets of genes from the 9,703 cDNA elements on the arrays. Before clustering and display, the logarithm of the measured fluorescence ratios for each gene were centred by subtracting the arithmetic mean of all ratios measured for that gene. The centring makes all subsequent analyses independent of the amount of each gene's mRNA in the reference pool.

We applied a hierarchical clustering algorithm separately to the cell lines and genes using the Pearson correlation coefficient as the measure of similarity and average linkage clustering[3,19-21]. The results of this process are two dendrograms (trees), one for the cell lines and one for the genes, in which very similar elements are connected by short branches, and longer branches join elements with diminishing degrees of similarity. For visual display the rows and columns in the initial data table were reordered to conform to the structures of the dendrograms obtained from the cluster analysis. Each cell in the cluster-ordered data table was replaced by a graded colour (pure red through black to pure green), representing the mean-adjusted ratio value in the cell. Gene labels in cluster diagrams are displayed here only for genes that were represented in the microarray by sequence-verified cDNAs. A complete software implementation of this process is available (http://rana.stanford.edu/software), as well as all clustering results (http://genome-www.stanford.edu/nci60).

1. Stinson, S.F. *et al.* Morphological and immunocytochemical characteristics of human tumor cell lines for use in a disease-oriented anticancer drug screen. *Anticancer Res.* 12, 1035–1053 (1992).
2. Myers, T.G. *et al.* A protein expression database for the molecular pharmacology of cancer. *Electrophoresis* 18, 647–653 (1997).
3. Weinstein, J.N. *et al.* An information-intensive approach to the molecular pharmacology of cancer. *Science* 275, 343–349 (1997).
4. Monks, A., Scudiero, D.A., Johnson, G.S., Paull, K.D. & Sausville, E.A. The NCI anti-cancer drug screen: a smart screen to identify effectors of novel targets. *Anticancer Drug Des.* 12, 533–541 (1997).
5. Paull, K.D. *et al.* Display and analysis of patterns of differential activity of drugs against human tumor cell lines: development of mean graph and COMPARE algorithm. *J. Natl Cancer Inst.* 81, 1088–1092 (1989).
6. Weinstein, J.N. *et al.* Neural computing in cancer drug development: predicting mechanism of action. *Science* 258, 447–451 (1992).
7. van Osdol, W.W., Myers, T.G., Paull, K.D., Kohn, K.W. & Weinstein, J.N. Use of the Kohonen self-organizing map to study the mechanisms of action of chemotherapeutic agents. *J. Natl Cancer Inst.* 86, 1853–1859 (1994).
8. DeRisi, J.L., Iyer, V.R. & Brown, P.O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686 (1997).
9. Iyer, V.R. *et al.* The transcriptional program in the response of human fibroblasts to serum. *Science* 283, 83–87 (1999).
10. Brown, P.O. & Botstein, D. Exploring the new world of the genome with DNA microarrays. *Nature Genet.* 21 (suppl.), 33–37 (1999).
11. Scherf, U. *et al.* A gene expression database for the molecular pharmacology of cancer. *Nature Genet.* 24, 236–244 (2000).
12. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* 58, 5009–5013 (1998).
13. Der, S.D., Zhou, A., Williams, B.R. & Silverman, R.H. Identification of genes differentially regulated by interferon-α, -β or -γ or using oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* 95, 15623–15628 (1998).
14. Alon, U. *et al.* Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl Acad. Sci. USA* 96, 6745–6750 (1999).
15. Wang, K. *et al.* Monitoring gene expression profile changes in ovarian carcinomas using cDNA microarray. *Gene* 229, 101–108 (1999).
16. Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* 96, 2907–2912 (1999).
17. Shalon, D., Smith, S.J. & Brown, P.O. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Res.* 6, 639–645 (1996).
18. Eisen, M.B. & Brown, P.O. DNA arrays for analysis of gene expression. *Methods Enzymol.* 303, 179–205 (1999).
19. Sokal, R.R. & Sneath, P.H.A. *Principles of Numerical Taxonomy* (W.H. Freeman, San Francisco, 1963).
20. Hartigan, J.A. *Clustering Algorithms* (Wiley, New York, 1975).
21. Eisen, M.B., Spellman, P.T., Brown, P.O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* 95, 14863–14868 (1998).
22. del Marmol, V. & Beermann, F. Tyrosinase and related proteins in mammalian pigmentation. *FEBS Lett.* 381, 165–168 (1996).
23. Kawakami, Y. *et al.* The use of melanosomal proteins in the immunotherapy of melanoma. *J. Immunother.* 21, 237–246 (1998).
24. Cailleau, R., Olive, M. & Cruciger, Q.V. Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization. *In Vitro* 14, 911–915 (1978).
25. Brinkley, B.R. *et al.* Variations in cell form and cytoskeleton in human breast carcinoma cells in vitro. *Cancer Res.* 40, 3118–3129 (1980).
26. Nesland, J.M., Holm, R., Johannessen, J.V. & Gould, V.E. Neuroendocrine differentiation in breast lesions. *Pathol. Res. Pract.* 183, 214–221 (1988).
27. Davies, J.A. & Garrod, D.R. Molecular aspects of the epithelial phenotype. *Bioessays* 19, 699–704 (1997).
28. Garrod, D., Chidgey, M. & North, A. Desmosomes: differentiation, development, dynamics and disease. *Curr. Opin. Cell Biol.* 8, 670–678 (1996).
29. Cowin, P. & Burke, B. Cytoskeleton-membrane interactions. *Curr. Opin. Cell Biol.* 8, 56–65 (1996); erratum: 8, 244 (1996).
30. Litvinov, S.V. *et al.* Epithelial cell adhesion molecule (Ep-CAM) modulates cell-cell interactions mediated by classic cadherins. *J. Cell Biol.* 139, 1337–1348 (1997).
31. Helmle-Kolb, C. *et al.* Na/H exchange activities in NHE1-transfected OK-cells: cell polarity and regulation. *Pflugers Arch.* 425, 34–40 (1993); erratum: 427, 387 (1994).
32. Manfruelli, P., Arquier, N., Hanratty, W.P. & Semeriva, M. The tumor suppressor gene, lethal(2)giant larvae (1(2)g1), is required for cell shape change of epithelial cells during Drosophila development. *Development* 122, 2283–2294 (1996).
33. Lincecum, J.M., Fannon, A., Song, K., Wang, Y. & Sassoon, D.A. Msh homeobox genes regulate cadherin-mediated cell adhesion and cell-cell sorting. *J. Cell Biochem.* 70, 22–28 (1998).
34. Hackett, A.J. *et al.* Two syngeneic cell lines from human breast tissue: the aneuploid mammary epithelial (Hs578T) and the diploid myoepithelial (Hs578Bst) cell lines. *J. Natl Cancer Inst.* 58, 1795–1806 (1977).
35. Rutka, J.T. *et al.* Establishment and characterization of a cell line from a human gliosarcoma. *Cancer Res.* 46, 5893–5902 (1986).
36. Nguyen, H., Hiscott, J. & Pitha, P.M. The growing family of interferon regulatory factors. *Cytokine Growth Factor Rev.* 8, 293–312 (1997).
37. Moscow, J.A., Schneider, E., Ivy, S.P. & Cowan, K.H. Multidrug resistance. *Cancer Chemother. Biol. Response Modif.* 17, 139–177 (1997).
38. Smith, H.S. & Hackett, A.J. The use of cultured human mammary epithelial cells in defining malignant progression. *Ann. N Y Acad. Sci.* 464, 288–300 (1986).
39. Rutka, J.T. *et al.* Establishment and characterization of five cell lines derived from human malignant gliomas. *Acta Neuropathol.* 75, 92–103 (1987).
40. Ronnov-Jessen, L., Petersen, O.W. & Bissell, M.J. Cellular changes involved in conversion of normal to malignant breast: importance of the stromal reaction. *Physiol. Rev.* 76, 69–125 (1996).
41. Perou, C.M. *et al.* Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. *Proc. Natl Acad. Sci. USA* 96, 9212–9217 (1999).
42. Bonner, R.F. *et al.* Laser capture microdissection: molecular analysis of tissue. *Science* 278, 1481–1483 (1997).
43. Sgroi, D.C. *et al.* In vivo gene expression profile analysis of human breast cancer progression. *Cancer Res.* 59, 5656–5661 (1999).

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68	A1	(11) International Publication Number: WO 95/21944
		(43) International Publication Date: 17 August 1995 (17.08.95)

(21) International Application Number: PCT/US95/01863

(22) International Filing Date: 14 February 1995 (14.02.95)

(30) Priority Data:
08/195,485   14 February 1994 (14.02.94)   US

(60) Parent Application or Grant
   (63) Related by Continuation
      US   08/195,485 (CIP)
      Filed on   14 February 1994 (14.02.94)

(71) Applicant *(for all designated States except US)*: SMITHKLINE BEECHAM CORPORATION [US/US]; Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: ROSENBERG, Martin [US/US]; 241 Mingo Road, Royersford, PA 19468 (US). DEBOUCK, Christine [BE/US]; 667 Pugh Road, Wayne, PA 19087 (US). BERGSMA, Derk [US/US]; 271 Irish Road, Berwyn, PA 19312 (US).

(74) Agents: JERVIS, Herbert, H. et al.; SmithKline Beecham Corporation, Corporate Intellectual Property, UW2220, 709 Swedeland Road, P.O. Box 1539, King of Prussia, PA 19406-0939 (US).

(81) Designated States: JP, US, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
   With international search report.

(54) Title: DIFFERENTIALLY EXPRESSED GENES IN HEALTHY AND DISEASED SUBJECTS

(57) Abstract

   The present invention involves methods and compositions for identifying genes which are differentially expressed in a normal healthy animal and an animal having a selected disease or infection, and methods for diagnosing diseases or infections characterized by the presence of those genes, despite the absence of knowledge about the gene or its function. The methods involve the use of a composition suitable for use in hybridization which consists of a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. Each sequence comprises a fragment of an EST isolated from an identified DNA library prepared from tissue or cell samples of a healthy animal, an animal with a selected disease or infection, and any combination thereof. Differences in hybridization patterns produced through use of this composition and the specified methods enable diagnosis of disease based on differential expression of genes of unknown function, and enable the identification of those genes and the proteins encoded thereby.

differentially expressed genes in healthy and diseased subjects

Cross Reference to Related Applications:

5   This application is a continuation-in-part application of U.S. Serial No.
08/195,485 filed February 14, 1994, the contents of which are incorporated herein by
reference.

Field of the Invention

10   The present invention relates to the use of immobilized
oligonucleotide/polynucleotide or polynucleotide sequences for the identification,
sequencing and characterization of genes which are implicated in disease, infection,
or development and the use of such identified genes and the proteins encoded thereby
in diagnosis, prognosis, therapy and drug discovery.

15

Background of the Invention

          Identification, sequencing and characterization of genes, especially
human genes, is a major goal of modern scientific research. By identifying genes,
determining their sequences and characterizing their biological function, it is possible
20   to employ recobinant DNA technology to produce large quantities of valuable "gene
products", e.g., proteins and peptides. Additionally, knowledge of gene sequences
can provide a key to diagnosis, prognosis and treatment of a variety of disease states
in plants and animals which are characterized by inappropriate expression and/or
repression of selected gene(s) or by the influence of external factors, e.g., carcinogens
25   or teratogens, on gene function. The term disease-associated genes(s) is used herein
in its broadest sence to mean not only genes associated with classical inherited
diseases, but also those associated with genetic predisposition to disease as well as
infectious or pathogenic states resulting from gene expression by infectious agents or
the effect on host cell gene expression by the presence of such a pathogen or its
30   products Locating disease-associated genes will permit the development of
diagnostic and prognostic reagents and methods, as well as possible therapeutic
regimens, and the discovery of new drugs for treating or preventing the occurrence of
such diseases.

          Methods have been described for the identification of certain novel
35   gene sequences, referred to as Expressed Sequence Tags (EST) [see, e.g., Adams et
al, Science, 252:1651-1656 (1991); and International Patent Application No.
WO93/00353, published January 7, 1993]. Conventially, an EST is a specific cDNA
polynucleotide sequence, or tag, about 150 to 400 nucleotides in length, derived from

a messenger RNA molecule by reverse transcription, which is a marker for, and component of, a human gene actually transcribed *in vivo*. However, as used herein an EST also refers to a genomic DNA fragment derived from an organism, such as a microorganism,the DNA of which lacks intron regions.

5   A variety of techniques have been described for identifying particular gene sequences on the basis of their gene products. For example, several techniques are described in the art [see, e.g., International Patent Application No. WO91/07087, published May 30, 1991]. Additionally, known methods exist for the amplification of desired sequences [see, e.g., International Patent Application No. WO91/17271,

10   published November 14, 1991, among others].

       However, at present, there exist no established methods for filling the need in the art for methods and reagents which employ fragments of differentially expressed genes of known, unknown (or previously unrecognized) function or consequence to provide diagnostic and therapeutic methods and reagents for diagnosis

15   and treatment of disease or infection, which conditions are characterized by such genes and gene products. It should be appreciated that it is the expression differences that are diagnostic of the altered state (e.g., predisease, disease, pathogenic, progression or infectious). Such genes associated with the altered state are likely to be the targets of drug discovery, whether the genes are the cause or the effect of the

20   condition, identification of such genes provides insight into which gene expression needs to be re-altered in order to reestablished the healthy state.

Summary of the Invention

       In one aspect, the invention provides methods for identifying gene(s)

25   which are differentially expressed, for example, in a normal healthy organism and an organism having a disease. The method involves producing and comparing hybridization patterns formed between samples of expressed mRNA or cDNA polynucleotide sequences obtained from either analogous cells, tissues or organs of a healthy organism and a diseased organism and a defined set of

30   oligonucleotide/polynucleotide/polynucleotide sequence probes from either an healthy organism or a diseased organism immobilized on a support. Those defined oligonucleotide/polynucleotide sequences are representative of the total expressed genetic component of the cells, tissues, organs or organism as defined the collection of partial cDNA sequences (ESTs). The differences between the hybridization

35   patterns permit identification of those particular EST or gene-specific oligonucleotide/polynucleotide sequences associated with differential expression, and the identification of the EST permits identification of the clone from which it was

derived and using ordinary skill further cloning and, if desired, sequencing of the full-length cDNA and genomic counterpart, i.e., gene, from which it was obtained.

In another aspect, the invention provides methods substantially similar to those described above, but which permit identification of those gene(s) of a
5   pathogen which are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected organism, hybridized to an oligonucleotide/polynucleotide set representative of the gene coding complement of the pathogen of interest.

In another aspect, the invention provides methods substantially similar
10   to those described above, but which permit identification of those ESTs-specific oligonucleotide/polynucleotide sequences of host gene(s) which represent genes being differentially expressed/ altered in expression by the disease state, or infection and are expressed in any biological sample of an infected organism based on comparative hybridization of RNA/cDNA samples derived from a healthy versus infected
15   organism of interest.

In a further aspect, the methods described above and in detail below, also provide methods for diagnosis of diseases or infections characterized by differentially expressed genes, the expression of which has been altered as a result of infection by the pathogen or disease causing agent in question. All identified
20   differences provide the basis for diagnostic testing be it the altered expression of endogenous genes or the patterned expression of the genes of the infecting organism. Such patterns of altered expression are defined by comparing RNA/cDNA from the two states hybridized against a panel of oligonucleotide/polynucleotides representing the expressed gene component of a cell, tissue, organ or organism as defined by its
25   collection of ESTs.

Yet a further aspect of this invention provides a composition suitable for use in hybridization, which comprises a solid surface on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence comprising a fragment of an EST isolated
30   from a cDNA or DNA library prepared from at least one selected tissue or cell sample of a healthy (i.e., pre-disease state) animal, at least one analogous sample of an animal having a disease, at least one analogous sample of an animal infected with a pathogen or the pathogen itself, or any combination or multiple combinations thereof.

An additional aspect of the invention provides an isolated gene
35   sequence which is differentially expressed in a normal healthy animal and an animal having a disease, and is identified by the methods above. Similarly, an isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal can be identified by the methods above.

3

Yet another aspect of the invention is that it provides not only a means for a static diagnostic but also provides a means for a carrying out the procedure over time to measure disease progression as well as monitoring the efficacy of disease treatment regimes including an toxicological effects thereof.

5   Another aspect of the invention is an isolated protein produced by expression of the gene sequences identified above. Such proteins are useful in therapeutic compositions or diagnostic compositions, or as targets for drug development.

Other aspects and advantages of the present invention are described

10   further in the following detailed description of the preferred embodiments thereof.

Detailed Description of the Invention

The present invention meets the unfulfilled needs in the art by providing methods for the identification and use of gene fragments and genes, even

15   those of unknown full length sequence and unknown function, which are differentially expressed in a healthy animal and in an animal having a specific disease or infection by use of ESTs derived from DNA libraries of healthy and/or diseased/infected animals. Employing the methods of this invention permits the resulting identification and isolation of such genes by using their corresponding ESTs

20   and thereby also permits the production of protein products encoded by such genes. The genes themselves and/or protein products, if desired, may be employed in the diagnosis or therapy of the disease or infection with which the genes are associated and in the development of new drugs therefor.

It has been appreciated that one or more differentially identified EST

25   or gene-specific oligonucleotide/polynucleotides define a pattern of differentially expressed genes diagnostic of a predisease, disease or infective state. A knowledge of the specific biological function of the EST is not required only that the ESTs identifies a gene or genes whose altered expression is associated reproducibly with the predisease, disease or infectious state. The differences permit the identification of

30   gene products altered in their expression by the disease and represent those products most likely to be targets of therapeutic intervention. Similarly, the product may be of the infecting organism itself and also be an effective target of intervention.

                  1.   Definitions.

35   Several words and phrases used throughout this specification are defined as follows:

As used herein, the term "gene" refers to the genomic nucleotide sequence from which a cDNA sequence is derived, which cDNA produces an EST, as

4

described below. The term gene classically refers to the genomic sequence, which, upon processing, can produce different cDNAs, e.g., by splicing events. However, for ease of reading, any full-length counterpart cDNA sequence which gives rise to an EST will also be referred to by shorthand herein as a 'gene'.

5   The term "organism" includes without limitation, microbes, plants and animals.

The term "animal" is used in its broadest sense to include all members of the animal kingdom, including humans. It should be understood, however, that according to this invention the same species of animal which provides the biological 10   sample also is the source of the defined immobilized oligonucleotide/polynucleotides as defined below.

The term "pathogen" is defined herein as any molecule or organism which is capable of infecting an animal or plant and replicating its nucleic acid sequences in the cells or tissues of that animal or plant . Such a pathogen is generally 15   associated with a disease condition in the infected animal or plant. Such pathogens may include viruses, which replicate intra- or extra-cellularly, or other organisms, such as bacteria, fungi or parasites; which generally infect tissues or the blood. Certain pathogens or microorganisms are known to exist in sequential and distinguishable stages of development, e.g., latent stages, infective stages, and stages 20   which cause symptomatic diseases. In these different stages, the pathogens are anticipated to express differentially certain genes and/or turn on or off host cell gene expression.

As used herein, the term "disease" or "disease state" refers to any condition which deviates from a normal or standardized healthy state in an organism 25   of the same species in terms of differential expression of the organism's genes. In other words, a disease state can be any illness or disorder be it of genetic or environmental origin , for example, an inherited disorder such as certain breast cancers, or a disorder which is characterized by expression of gene(s) normally in an inactive, 'turned off' state in a healthy animal, or a disorder which is characterized by 30   under-expression or no expression of gene(s) which is normally activated or 'turned on' in a normal healthy animal. Such differential expression of genes may also be detected in a condition caused by infection, inflammation, or allergy, a condition caused by development or aging of the animal, a condition caused by administration of a drug or exposure of the animal to another agent, e.g., nutrition, which affects 35   gene expression. Essentially, the methods described herein can be adapted to detect differential gene expression resulting from any cause, by manipulation of the defined oligonucleotide/polynucleotides and the samples tested as described below. The

concept of disease or disease state also includes its temporal aspects in terms of progression and treatment.

The phrase "differentially expressed" refers to those situations in which a gene transcript is found in differing numbers of copies, or in activated vs

5   inactivated states, in different cell types or tissue types of an organism, having a selected disease as contrasted to the levels of the gene transcript found in the same cells or tissues of a healthy organism. Genes may be differentially expressed in differing states of activation in microorganisms or pathogens in different stages of development. For example, multiple copies of gene transcripts may be found in an

10   organism having a selected disease, while only one, or significantly fewer copies, of the same gene transcript are found in a healthy organism, or vice-versa.

As used herein, the term "solid support" refers to any known substrate which   is   useful   for   the   immobilization   of   large   numbers   of oligonucleotide/polynucleotide sequences by any available method to enable

15   detectable hybridization of the immobilized oligonucleotide/polynucleotide sequences with other polynucleotide sequences in a sample. Among a number of available solid supports, one desirable example is the supports described in International Patent Application No. WO91/07087, published May 30, 1991.Also useful are suports such as but not limited to nitrocellulose, mylein, glass, silica ans Pall Biodyne C® It is

20   also anticipated that improvements yet to be made to conventional solid supports may also be employed in this invention.

The term "surface" means any generally two-dimensional structure on a solid support to which the desired oligonucleotide/polynucleotide sequence is attached or immobilized. A surface may have steps, ridges, kinks, terraces and the

25   like.

As used herein, the term "predefined region" refers to a localized area on a surface of a solid support on which is immobilized one or multiple copies of a particular oligonucleotide/polynucleotide sequence and which enables the identification of the oligonucleotide/polynucleotide at the position, if hybridization of

30   that oligonucleotide/polynucleotide to a sample polynucleotide occurs.

By   "immobilized"   refers   to   the   attachment   of   the oligonucleotide/polynucleotide to the solid support. Means of immobilization are known and conventional to those of skill in the art, and may depend on the type of support being used.

35   By "EST" or "Expressed Sequence Tag" is meant a partial DNA or cDNA sequence of about 150 to 500, more preferably about 300, sequential nucleotides of a longer sequence obtained from a genomic or cDNA library prepared from a selected cell, cell type, tissue or tissue type, organ or organism which longer

6

sequence corresponds to an mRNA of a gene found in that library. An EST is generally DNA. One or more libraries made from a single tissue type typically provide at least about 3000 different (i.e., unique) ESTs and potentially the full complement of all possible ESTs representing all cDNAs e.g., 50,000-100,000 in an
5   animal such as a human. Further background and information on the construction of ESTs is described in M. D. Adams et al, Science, 252:1651-1656 (1991); and International Application Number PCT/US92/05222 (January 7, 1993).

As used herein, the term "defined oligonucleotide/polynucleotide sequence" refers to a known nucleotide sequence fragment of a selected EST or gene.
10   This term is used interchangeably with the term "fragments of EST". These sequential sequences are generally comprised of between about 15 to about 45 nucleotides and more preferably between about 20 to about 25 nucleotides in length. Thus any single EST of 300 nucleotides in length may provide about 280 different defined oligonucleotide/polynucleotide sequences of 20 nucleotides in length (e.g.,
15   20-mers). The lengths of the defined oligonucleotide/polynucleotides may be readily increased or decreased as desired or needed, depending on the limitations of the solid support on which they may be immobilized or the requirements of the hybridization conditions to be employed.The length is generally guided by the principle that it should be of sufficient length to insure that it is one average only represented once in
20   the   population   to   be   examined.   Generally,   these   defined oligonucleotide/polynucleotides are RNA or DNA and are preferably derived from the anti-sense strand of the EST sequence or from a corresponding mRNA sequence to enable their hybridization with samples of RNA or DNA. Modified nucleotides may be incorporated to increase stability and hybridization properties.
25   By the term "plurality of defined oligonucleotide/polynucleotide sequences" is meant the following. A surface of a solid support may immobilize a large number of "defined oligonucleotide/polynucleotides". For example, depending upon the nature of the surface, it can immobilize from about 300 to upwards of 60,000 defined 20-mer oligonucleotide/polynucleotides. It is anticipated that future
30   improvements to solid surfaces will permit considerably larger such pluralities to be immobilized on a single surface. A "plurality" of sequences refers to the use on any one solid support of multiple different defined oligonucleotide/polynucleotides from a single EST from a selected library, as well as multiple different defined oligonucleotide/polynucleotides from different ESTs from the same library or many
35   libraries from the same or different tissues, and may also include multiple identical copies of defined oligonucleotide/polynucleotides. Ultimately a pluarality has at least one oligonucleotide/polynucleotide per expressed gene in the entire organism For example, from a library producing about 5,000-10,000 ESTs, a single support can

include at least about 1-20 defined oligonucleotide/polynucleotides representing every EST in that library. The composition of defined oligonucleotide/polynucleotides which make up a surface according to this invention may be selected or designed as desired.

5   The term "sample" is employed in the description of this invention in several important ways. As used herein, the term "sample" encompasses any cell or tissue from an organism. Any desired cell or tissue type in any desired state may be selected to form a sample. For example, the sample cell desired may be a human T cell; the desired cell type for use in this invention may be a quiescent T cell or an

10   activated T cell.

By the phrase "analogous sample" or "analogous cell or tissue" is meant that according to this invention when the ESTs which provide the defined oligonucleotide/polynucleotides are produced from a cDNA library prepared from a single tissue or cell type source sample, e.g., liver tissue of a human, then the samples

15   used to hybridize to those immobilized defined oligonucleotide/polynucleotides are preferably provided by the same type of sample from either a healthy or diseased animal, i.e., liver tissue of a healthy human and liver tissue of a diseased or infected human or from a human suspected of having that disease or infection. Alternatively, if the surface contains defined oligonucleotide/polynucleotides from multiple cells or

20   tissues, then the "samples" which are hybridized thereto can be but are not limited to samples obtained from analogous multiple tissues or cells.

By the term "detectably hybridizing" means that the sample from the healthy organism or diseased or infected organism is contacted with the defined oligonucleotide/polynucleotides on the surface for sufficient time to permit the

25   formation of patterns of hybridization on the surfaces caused by hybridization between certain polynucleotide sequences in the samples with the certain immobilized defined oligonucleotide/polynucleotides. These patterns are made detectable by the use of available conventional techniques, such as fluorescent labelling of the samples. Preferably hybridization takes place under stringent conditions, e.g., revealing

30   homologies of about 95%. However, if desired, other less stringent conditions may be selected. Techniques and conditions for hybridization at selected stringencies are well known in the art [see, e.g., Sambrook et al, Molecular Cloning. A Laboratory Manual., Cold Spring Harbor Laboratory, Cold Spring Harbor, NY (1989)].

35   II.   Compositions of The Invention
The present invention is based upon the use of ESTs from any desired cell or tissue in known technologies for oligonucleotide/polynucleotide hybridization.

8

A.   *ESTs*

An EST, as defined above, is for an animal, a sequence from a cDNA clone that corresponds to an mRNA. The EST sequences useful in the present invention are isolated preferably from cDNA libraries using a rapid screening and

5   sequencing technique.   Custom made cDNA libraries are made using known techniques. See, generally, Sambrook et al, cited above. Briefly, mRNA from a selected cell or tissue is reverse transcribed into complementary DNA (cDNA) using the reverse transcriptase enzyme and made double-stranded using RNase H coupled with DNA polymerase or reverse transcriptase. Restriction enzyme sites are added to

10   the cDNA and it is cloned into a vector. The result is a cDNA library. Alternatively, commercially available cDNA libraries may be used. Libraries of cDNA can also be generated from recombinant expression of genomic DNA using known techniques, including polymerase chain reaction-derived techniques.

ESTs (which can range from about 150 to about 500 nucleotides in

15   length, preferably about 300 nucleotides) can be obtained through sequence analysis from either end of the cDNA insert. Desirably, the DNA libraries used to obtain ESTs use directional cloning methods so that either the 5' end of the cDNA (likely to contain coding sequence) or the 3' end (likely to be a non-coding sequence) can be selectively obtained.

20   In general, the method for obtaining ESTs comprises applying conventional   automated   DNA   sequencing   technology   to   screen   clones, advantageously randomly selected clones, from a cDNA library. The cDNA libraries from the desired tissue can be preprocessed, or edited, by conventional techniques to reduce repeated sequencing of high and intermediate abundance clones and to

25   maximize the chances of finding rare messages from specific cell populations. Preferably, preprocessing includes the use of defined composition prescreening probes, e.g., cDNA corresponding to mitochondria, abundant sequences, ribosomes, actins, myelin basic polypeptides, or any other known high abundance peptide. These prescreening probes used for preprocessing are generally derived from known ESTs.

30   Other   useful   preprocessing   techniques   include   subtraction   hybridization,   which preferentially reduces the population of highly represented sequences in the library [e.g., see Fargnoli et al, Anal. Biochem., 187:364 (1990)] and normalization, which results in all sequences being represented in approximately equal proportions in the library [Patanjali et al, Proc. Natl. Acad. Sci. USA, 88:1943 (1991)]. Additional

35   prescreening/differential screening approaches are known to those skilled in the art.

ESTs can then be generated from partial DNA sequencing of the selected clones. The ESTs useful in the present invention are preferably generated using low redundancy of sequencing, typically a single sequencing reaction. While

9

single sequencing reactions may have an accuracy as low as 90%, this nevertheless provides sufficient fidelity for identification of the sequence and design of PCR primers.

If desired, the location of an EST in a full length cDNA is determined
5   by analyzing the EST for the presence of coding sequence. A conventional computer program is used to predict the extent and orientation of the coding region of a sequence (using all six reading frames). Based on this information, it is possible to infer the presence of start or stop codons within a sequence and whether the sequence is completely coding or completely non-coding or a combination of the two. If start
10   or stop codons are present, then the EST can cover both part of the 5'-untranslated or 3'-untranslated part of the mRNA (respectively) as well as part of the coding sequence. If no coding sequence is present, it is likely that the EST is derived from the 3' untranslated sequence due to its longer length and the fact that most cDNA library construction methods are biased toward the 3' end of the mRNA. It should be
15   understood that both coding and non-coding regions may provide ESTs equally useful in the described invention.

A number of specific ESTs suitable for use in the present invention are described above Adams et al (supra), which may be incorporated by reference herein, to describe non-essential examples of desirable ESTs. Other ESTs
20   exist in the art which may also be useful in this invention, as will ESTs yet to be developed by these known techniques.

B.   Preparing the Solid Support of the Invention

Oligonucleotide sequences which are fragments of defined sequence are derived from each EST by conventional means, e.g., conventional
25   chemical   synthesis   or   recombinant   techniques.   Each   defined oligonucleotide/polynucleotide sequence as described above is a fragment, can be, but is not necessarily an anti-sense fragment, of an EST isolated from a DNA library prepared from a selected cell or tissue type from a selected animal. For use in the present   invention,   it   is   presently   preferred   that   the   defined
30   oligonucleotide/polynucleotide sequences are 20-25mers. As described above, for each EST a number of such 20-25mers may be generated. The lengths may vary as described   above   as   well   as   the   composition.   For   example oligonucleotide/polynucleotides can be modified based on the Oligo 4.0 or simiolar programs to predict hybridization potential or to include modifieid nucleotides for the
35   reasons given above. It is alos appreciated that large DNA segments may be employed including entire ESTs or even full length genes particular when inserted into cloning vectors.

10

A plurality of these defined oligonucleotide/polynucleotide sequences are then attached to a selected solid support conventionally used for the attachment of nucleotide sequences again by known means. In contrast to other technologies available in the art, this support is designed to contain defined, not

5   random, oligonucleotide/polynucleotide sequences. The EST fragments, or defined oligonucleotide/polynucleotide sequences, immobilized on the solid support can include fragments of one or more ESTs from a library of at least one selected tissue or cell sample of a healthy animal, at least one analogous sample of the animal having a disease, at least one analogous sample of the animal infected with a pathogen, and

10   any combination thereof.

Numerous conventional methods are employed for attaching biological molecules such as oligonucleotide/polynucleotide sequences to surfaces of a variety of solid supports. See, e.g., Affinity Techniques, Enzyme Purification: Part B, Methods in Enzymology, Vol. 34, ed. W.B. Jakoby, M. Wilcheck, Acad. Press,

15   NY (1974); Immobilized Biochemicals and Affinity Chromatography, Advances in Experimental Medicine and Biology, vol. 42, ed. R. Dunlap, Plenum Press, NY (1974); U. S. Patent No. 4,762,881; U. S. Patent No. 4,542,102; European Patent Publication No. 391,608 (October 10, 1990); U. S. Patent No. 4,992,127 (Nov. 21, 1989).

20   One desirable method for attaching oligonucleotide/polynucleotide sequences derived from ESTs to a solid support is described in International Application No. PCT/US90/06607 (published May 30, 1991). Briefly, this method involves forming predefined regions on a surface of a solid support, where the predefined regions are capable of immobilizing ESTs. The

25   methods make use of binding substances attached to the surface which enable selective activation of the predefined regions. Upon activation, these binding substances become capable of binding and immobilizing oligonucleotide/polynucleotides based on EST or longer gene sequences.

Any of the known solid substrates suitable for binding

30   oligonucleotide/polynucleotides at pre-defined regions on the surface thereof for hybridization and methods for attaching the oligonucleotide/polynucleotides thereto may be employed by one of skill in the art according to this invention. Similarly, known conventional methods for making hybridization of the immobilized oligonucleotide/polynucleotides detectable, e.g., fluorescence, radioactivity,

35   photoactivation, biotinylation, solid state circuitry, and the like may be used in this invention.

Thus, by resorting to known techniques, the invention provides a composition suitable for use in hybridization which consists of a surface of a solid

11

support on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization. For example, one composition of this invention is a solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type, e.g., a human

5   stem cell, or a single tissue, e.g., human liver, from a healthy human. Still another composition of this invention is another solid support on which are immobilized oligos of EST fragments from a library constructed from a single cell type or a tissue from a human having a selected disease or predispositon to a selected disease, e.g., liver cancer.

10   Another embodiment of the compositions of this invention include a single solid support having oligonucleotides of ESTs from both single cell or single tissue libraries from both a healthy and diseased human. Still other embodiments include a single support on which are immobilized oligos of EST fragments from more than one tissue or cell library from a healthy human or a single

15   support on which are immobilized more than one tissue or cell library from both healthy and diseased animals or humans. A preferred composition of this invention is anticipated to be a single support containing oligos of ESTs for all known cells and tissues from a selected organism.

20   *III.   The Methods of the Invention*
                   A.   Identification of Genes
                   The present invention employs the compositions described above in methods for identifying genes which are differentially expressed in a normal healthy organism and an organism having a disease or infection. These methods may

25   be employed to detect such genes, regardless of the state of knowledge about the function of the gene. The method of this invention by use of the compositions containing multiple defined EST fragments from a single gene as described above is able to detect levels of expression of genes or in other cases simply the expression or lack thereof, which differ between normal, healthy organisms and organisms having a

30   selected disease, disorder or infection.
                   One such method employs a first surface of a solid support on which is immobilized at pre-defined regions thereon a plurality of defined oligonucleotide/polynucleotide sequences, described above, of ESTor longer gene fragment isolated from a cDNA library prepared from at least one selected tissue or

35   cell sample of a healthy animal (the "healthy test surface") and a second such surface on which is immobilized at pre-defined regions a plurality of defined oligonucleotide/polynucleotide sequences of ESTor longer gene fragment isolated from at least one analogous tissue of an animal having a selected disease (the "disease

                                          12

test surface"). These test surfaces may be standardized for the selected animal or selected cell or tissue sample from that animal (i.e., they are prescreened for polymorphisms in the species population).

Polynucleotide sequences are then isolated from mRNA and/or cDNA from a biological sample from a known healthy animal ("healthy control") and a second sample is similarly prepared from a sample from a known diseased animal ("disease sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides.

According to the method the healthy control sample is contacted with one set of the healthy test surface and the disease test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the disease test surface.

In a similar manner, the disease sample is detectably hybridized to another set of healthy test and disease test surfaces, forming a third hybridization pattern between the disease sample and healthy test surface and a fourth hybridization pattern between the disease sample and the disease test surface.

Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTor longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

In another embodiment of the method of this invention, the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample and the disease test sample surfaces are immobilized on a single solid support. For example, each fragment of an EST or longer gene fragment on the surface is isolated from at least two cDNA libraries prepared from a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having a disease.

According to this embodiment, the healthy control sample is detectably hybridized to a copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal. Similarly, the disease sample is detectably hybridized to a second

13

copy of this single solid surface, forming one hybridization pattern with oligonucleotide/polynucleotides associated with both the healthy and diseased animal.

Comparing the two hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed

5   between the healthy control and the disease sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding EST or longer gene fragment from which the oligonucleotide/polynucleotides are obtained.

10   The identification of one or more ESTs as the source of the defined oligonucleotide/polynucleotide which produced a "difference" in hybridization patterns according to these methods permits ready identification of the gene from which those ESTs were derived. Because oligonuleotides are of sufficient length that they will hybridize under stringent conditions only with a RNA/cDNA for

15   that gene to which they correspond, the oligo can be used to identify the EST and in turn the clone from which it was derived and by subsequent cloning, obtain the sequence of the full-length cDNA and its genomic counterparts, i.e., the gene, from which it was obtained.

In other words, the ESTs identified by the method of this

20   invention can be employed to determine the complete sequence of the mRNA, in the form of transcribed cDNA, by using the EST as a probe to identify a cDNA clone corresponding to a full-length transcript, followed by sequencing of that clone. The EST or the full length cDNA clone can also be used as a probe to identify a genomic clone or clones that contain the complete gene including regulatory and promoter

25   regions, exons, and introns.

It should be appreciated that one does not have to be restricted in using ESTs from a particular tissue from which probe RNA or cDNA is obtained, rather any or all ESTs (known or unknown) may be placed on the support. Hybridization will be used a form diagnostic patterns or to identifiy which particular

30   EST is detected. For example, all known ESTs from an organism are used to produce a "master" solid support to which control sample and disease samples are alternately hybridized. One then detects a pattern of hybridization associated with the particular disaease state which then forms the basis of a diagnostic test or the isolation of disease specific ESTs from which the intact gene may be cloned and sequenced

35   leading uiltimately to a defined therapuetic target.

Methods for obtaining complete gene sequences from ESTs are well-known to those of skill in the art. See, generally, Sambrook et al, cited above. Briefly, one suitable method involves purifying the DNA from the clone that was

14

sequenced to give the EST and labeling the isolated insert DNA. Suitable labeling systems are well known to those of skill in the art [see, eg. Basic Methods in Molecular Biology, L. G. Davis et al, ed., Elsevier Press, NY (1986)]. The labeled EST insert is then used as a probe to screen a lambda phage cDNA library or a

5 plasmid cDNA library, identifying colonies containing clones related to the probe cDNA which can be purified by known methods. The ends of the newly purified clones are then sequenced to identify full length sequences and complete sequencing of full length clones is performed by enzymatic digestion or primer walking. A similar screening and clone selection approach can be applied to clones from a

10 genomic DNA library.

Additionally, an EST or gene identified by this method as associated with inherited disorders can be used to determine at what stage during embryonic development the selected gene from which it is derived is developed by screening embryonic DNA libraries from various stages of development, e.g. 2-cell,

15 8-cell, etc., for the selected gene. As has been mentioned above, the invention may be applied in addtional temporal modes for monitoring the progression of a disease state, the efficacy of a particular treatment modality or the aging process of an individual.

Thus, the methods of this invention permit the identification,

20 isolation and sequencing of a gene which is differentially expressed in a selected disease/infection. As described in more detail below, the identified gene may then be employed to obtain any protein encoded thereby, or may be employed as a target for diagnostic methods or therapeutic approaches to the treatment of the disease, including, e.g., drug development.

25 The same methods as described above for the identification of genes, including genes of unknown function, which are differentially expressed in a disease state, may also be employed to identify other genes of interest. For example, another embodiment of this invention includes a method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with that

30 pathogen or the gene of the host which is altered in its expression as a result of the infection.

One such method employs a healthy test surface as described above, employing defined oligonucleotide/polynucleotides from a sample of a healthy, uninfected animal. The second such surface has immobilized at pre-defined

35 regions thereon a plurality of defined oligonucleotide/polynucleotide sequences of ESTs isolated from at least one analogous tissue or cell sample of an infected animal (the "infection test surface"). Polynucleotide sequences are isolated from a biological sample from a healthy animal ("healthy control") and a second sample is similarly

15

prepared from an animal infected with the selected pathogen ("infection sample"). These two samples are desirably selected from the cell or tissue analogous to that which provided the immobilized oligonucleotide/polynucleotides. It would also be possible to provide samples from the nucleic acid of the pathogen itself.

5   According to the method the healthy control sample is contacted with one set of the healthy test surface and the infection test surface described above for a time sufficient to permit detectable hybridization to occur between the sample and the immobilized defined oligonucleotide/polynucleotides on each surface. The results of this hybridization are a first hybridization pattern formed

10   between the nucleotides of healthy control and the healthy test surface and a second hybridization pattern formed between the nucleotides of healthy control sample and the infection test surface.

          In a similar manner, the infection sample is detectably hybridized to another set of healthy test and infection test surfaces, forming a third

15   hybridization pattern between the infection sample and healthy test surface and a fourth hybridization pattern between the infection sample and the infection test surface.

          Comparing the four hybridization patterns permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed

20   between the healthy animal and the animal infected with the pathogen by the presence of differences in the hybridization patterns at pre-defined regions. As mentioned differential expression is not required and simple qualitative analysis is possible by reference to gene expression which is simply present or absent.

          A second embodiment of this method parallels the second

25   embodiment of the method as applied to disease above, i.e., the same process is employed, with the exception that plurality of defined oligonucleotide/polynucleotide sequences forming the healthy test sample surface and the infection test sample surface are immobilized on a single solid support. The resulting first hybridization pattern (healthy control sample with healthy/infection test sample) and second

30   hybridization pattern (infection sample with healthy/infection test sample) permits detection of those defined oligonucleotide/polynucleotides which are differentially expressed between the healthy control and the infection sample by the presence of differences in the hybridization patterns at pre-defined regions. The oligonucleotide/polynucleotides on each surface which correspond to the pattern

35   differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained.

          As described above for the methods for identifying differential gene expression between diseased and healthy animals, the

16

oligonucleotide/polynucleotides on each surface which correspond to the pattern differences may be readily identified with the corresponding ESTs from which the oligonucleotide/polynucleotide sequences are obtained and the genes expressed by the pathogen identified for similar purposes. Other embodiments of these methods may

5   be developed with resort to the teaching herein, by altering the samples which provide the defined oligonucleotide/polynucleotides. For example, an EST, identified with a differentially expressed gene by the method of this invention is also useful in detecting genes expressed in the various stages of an pathogen's development, particularly the infective stage and following the cours of drug treatment and

10   emergence of resistant variants. For example, employing the techniques described above, the EST can be used for detecting a gene in various stages of the parasitic *Plasmodium* species life cycle, which include blood stages, liver stages, and gametocyte stages.

B.   *Diagnostic Methods*

15   In addition to use of the methods and compositions of this invention for identifying differentially expressed genes, another embodiment of this invention provides diagnostic methods for diagnosing a selected disease state, or a selected state resulting from aging, exposure to drugs or infection in an animal. According to this aspect of the invention, a first surface, described as the healthy test

20   surface above, and a second surface, described as the disease test surface or infection test surface, are prepared depending on the disease or infection to be diagnosed. The same processes of detectable hybridization to a first and second set of these surfaces with the healthy control sample and disease/infection sample are followed to provide the four above-described hybridization patterns, i.e., healthy control sample with

25   healthy test surface; healthy control sample with disease/infection test surface; disease/infection sample with healthy test surface; and disease/infection sample with disease/infection test surface.

The diagnosis of disease or infection is provided by comparing the four hybridization patterns. Substantial differences between the first and third

30   hybridization patterns, respectively, and the second and fourth hybridization patterns, respectively, indicate the presence of the selected disease or infection in said animal. Substantial similarities in the first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

A similar embodiment utilizes the single surface bearing both

35   the healthy test surface defined oligonucleotide/polynucleotides and the disease/infection test surface defined oligonucleotide/polynucleotides as described above. Parallel process steps as described above for detection of genes differentially expressed in disease and infected states are followed, resulting in a first hybridization

17

pattern (healthy control sample with single healthy and disease/infection test sample) and a second hybridization pattern (disease/infection sample with another copy of the single healthy and disease/infection test sample).

Diagnosis is accomplished by comparing the two hybridization

5   patterns, wherein substantial differences between the first and second hybridization patterns indicate the presence of the selected disease or infection in the animal being tested. Substantially similar first and second hybridization patterns indicate the absence of disease or infection. This like many of the foregoing embodiments may use known or unknown ESTs derived from many libraries.

10   *C.   Other Methods of the Invention*

As is obvious to one of skill in the art upon reading this disclosure, the compositions and methods of this invention may also be used for other similar purposes. For example, the general methods and compositions may be adapted easily by manipulation of the samples selected to provide the standardized

15   defined oligonucleotide/polynucleotides, and selection of the samples selected for hybridization thereto. One such modification is the use of this invention to identify cell markers of any type, e.g., markers of cancer cells, stem cell markers, and the like. Another modification involves the use of the method and compositions to generate hybridization patterns useful for forensic identification or an 'expression fingerprint'

20   of genes for identification of one member of a species from another. Similarly, the methods of this invention may be adapted for use in tissue matching for transplantation purposes as well as for molecular histology, i.e., to enable diagnosis of disease or disorders in pathology tissue samples such as biopsies. Still another use of this method is in monitoring the effects of development and aging upon the gene

25   expression in a selected animal, by preparing surfaces bearing oligonucleotide/polynucleotides prepared from samples of standardized younger members of the species being tested. Additionally the patient can serve as an internal control by virtue of having the method applied to blood samples every 5-10 years during his lifetime.

30   Still another intriguing use of this method is in the area of monitoring the effects of drugs on gene expression, both in laboratories and during clinical trials with animal, especially humans. Because the method can be readily adapted by altering the above parameters, it can essentially be employed to identify differentially expressed genes of any organism, at any stage of development, and

35   under the influence of any factor which can affect gene expression.

18

IV.   The Genes and Proteins Identified

Application of the compositions and methods of this invention as above described also provide other compositions, such as any isolated gene sequence which is differentially expressed between a normal healthy animal and an animal

5   having a disease or infection. Another embodiment of this invention is any isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal. Similarly an embodiment of this invention is any gene sequence identified by the methods described herein.

These gene sequences may be employed in conventional methods to

10   produce isolated proteins encoded thereby. To produce a protein of this invention, the DNA sequences of a desired gene identified by the use of the methods of this invention or portions thereof are inserted into a suitable expression system. Desirably, a recombinant molecule or vector is constructed in which the polynucleotide sequence encoding the protein is operably linked to a heterologous

15   expression control sequence permitting expression of the human protein. Numerous types of appropriate expression vectors and host cell systems are known in the art for mammalian (including human) expression, insect, e.g., baculovirus expression, yeast, fungal, and bacterial expression, by standard molecular biology techniques.

The transfection of these vectors into appropriate host cells, whether

20   mammalian, bacterial, fungal, or insect, or into appropriate viruses, can result in expression of the selected proteins. Suitable host cells or cell lines for transfection, and viruses, as well as methods for the construction and transfection of such host cells and viruses are well-known. Suitable methods for transfection, culture, amplification, screening, and product production and purification are also known in the art.

25   The genes and proteins identified by this invention can be employed, if desired in diagnostic compositions useful for the diagnosis of a disease or infection using conventional diagnostic assays. For example, a diagnostic reagent can be developed which detectably targets a gene sequence or protein of this invention in a biological sample of an animal. Such a reagent may be a complementary nucleotide

30   sequence, an antibody (monoclonal, recombinant or polyclonal), or a chemically derived agonist or antagonist.   Alternatively, the proteins and polynucleotide sequences of this invention, fragments of same, or complementary sequences thereto, may themselves be useful as diagnostic reagents for diagnosing disease states with which the ESTs of the invention are associated. These reagents may optionally be

35   labelled using diagnostic labels, such as radioactive labels, colorimetric enzyme label systems and the like conventionally used in diagnostic or therapeutic methods, e.g, Northern and Western blotting, antigen-antibody binding and the like. The selection of the appropriate assay format and label system is within the skill of the art and may

19

readily be chosen without requiring additional explanation by resort to the wealth of art in the diagnostic area.

Additionally, genes and proteins identified according to this invention may be used therapeutically. For example, the EST-containing gene sequences may be useful in gene therapy, to provide a gene sequence which in a disease is not properly or sufficiently expressed. In such a method, a selected gene sequence of this invention is introduced into a suitable vector or other delivery system for delivery to a cell containing a defect in the selected gene. Suitable delivery systems are well known to those of skill in the art and enable the desired EST or gene to be incorporated into the target cell and to be translated by the cell. The EST or gene sequence may be introduced to mutate the existing gene by recombination or provide an active copy thereof in addition to the inactive gene to replace its function.

Alternatively, a protein encoded by an EST or gene of the invention may be useful as a therapeutic reagent for delivery of a biologically active protein, particularly when the disease state is associated with a deficiency of this protein. Such a protein may be incorporated into an appropriate therapeutic formulation, alone or in combination with other active ingredients. Methods of formulating such therapeutic compositions, as well as suitable pharmaceutical carriers, and the like, are well known to those of skill in the art. Still an additional method of delivering the missing protein encoded by an EST, or the gene from which a selected EST was derived, involves expressing it directly *in vivo*. Systems for such *in vivo* expression are well known in the art.

Yet another use of the ESTs, genes identified according to the methods of this invention, or the proteins encoded thereby is a target for the screening and development of natural or synthetic chemical compounds which have utility as therapeutic drugs for the treatment of disease states associated with the identified genes and ESTs derived therefrom. As one example, a compound capable of binding to such a protein encoded by such a gene and either preventing or enhancing its biological activity may be a useful drug component for the treatment or prevention of such disease states.

Conventional assays and techniques may be used for the screening and development of such drugs. As one example, a method for identifying compounds which specifically bind to or inhibit or activate proteins encoded by these gene sequences can include simply the steps of contacting a selected protein or gene product, with a test compound to permit binding of the test compound to the protein; and determining the amount of test compound, if any, which is bound to the protein. Such a method may involve the incubation of the test compound and the protein immobilized on a solid support. Still other conventional methods of drug screening

can involve employing a suitable computer program to determine compounds having similar or complementary chemical structures to that of the gene product or portions thereof and screening those compounds either for competitive binding to the protein to detect enhanced or decreased activity in the presence of the selected compound.

5   Thus, through use of such methods, the present invention is anticipated to provide compounds capable of interacting with these genes, ESTs, or encoded proteins, or fragments thereof, and either enhancing or decreasing the biological activity, as desired. Such compounds are believed to be encompassed by this invention.

10   Numerous modifications and variations of the present invention are included in the above-identified specification and are expected to be obvious to one of skill in the art. Such modifications and alterations to the compositions and processes of the present invention are believed to be encompassed in the scope of the claims appended hereto.

15

WHAT IS CLAIMED IS:

1.   A method for identifying genes which are differentially expressed in
two different pre-determined states of an organism comprising:

5   a. providing a first surface on which is immobilized at pre-defined
regions on said surface a plurality of defined oligonucleotide/polynucleotide
. sequences, each sequence selected from the group consisting of a fragment of an EST,
an entire EST a fragment of a gene or an entire gene, isolated from a DNA library
prepared from at least one selected cell, tissue, organ or organism sample in a first
10   state and present in excess relative to the polynucleotide to be hybridized;

               b. providing a second surface on which is immobilized at pre-defined
regions on said surface a plurality of defined oligonucleotide/polynucleotide
sequences, each sequence selected from the group consisting of a fragment of an EST,
an entire EST a fragment of a gene or an entire gene, isolated from a DNA library
15   prepared from at least one selected cell, tissue, organ or organism sample in a second
state and present in excess relative to the polynucleotide to be hybridized;

               c. detectably hybridizing to a set of said first and second surfaces
polynucleotide sequences isolated from a sample from a said organism in said first
state, said sample selected from sources analogous to the sources of step (a), said
20   hybridization sufficient to form a first and second hybridization pattern on each said
first and second surface,

               d. detectably hybridizing to a set of said first and second surfaces
polynucleotide sequences isolated from a sample from said organism in said second
state, said sample selected from sources analogous to the sources of step (c), said
25   hybridization sufficient to form a third and fourth hybridization pattern on each said
first and second surface,

               e.   comparing at least two of the four hybridization patterns,
wherein genes differentially expressed in said first and second states are identified by
the presence of differences in the hybridization patterns at pre-defined regions;

30   f.   identifying the oligonucleotide/polynucleotides on each surface
which correspond to said pattern differences and the corresponding ESTs or larger
gene fragment from which the oligonucleotide/polynucleotides were obtained,
whereby identification of the EST or larger gene fragment permits identification of
the gene from which the ESTs or larger gene fragment were derived.

35

2. The method according to Claim 1 wherein said first and second states are respectively healthy and disease; pathogen uninfected and pathogen infected; a first progression state and a second progression of a disease or infection; a first treatment state and a second treatment state of a disease or infection; or a first developmental

5   and a second developmental state.

3. The method according to Claim 1 wherein said organism is a plant or an animal.

10   4. The method according to Claim 3 wherein said aniaml is a human.

5.   A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a.   providing a first surface on which is immobilized at pre-
15   defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample in a healthy animal and present in excess relative to the polynucleotide to be
20   hybridized;

b.   providing a second surface on which is immobilized at pre-
defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene, isolated
25   from a DNA library prepared from at least one selected cell, tissue, organ or organism sample from an animal having said disease and present in excess relative to the polynucleotide to be hybridized;

c.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample
30   selected from sources analogous to the sources of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

d.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each
5   said first and second surface,

e.   comparing at least two of the four hybridization patterns, wherein genes differentially expressed in said first and second states are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f.   identifying the oligonucleotide/polynucleotides on each surface
10   which correspond to said pattern differences and the corresponding ESTs or larger gene fragment from which the oligonucleotide/polynucleotides were obtained, whereby identification of the EST or larger gene fragment permits identification of the gene from which the ESTs or larger gene fragment were derived.

15   6.   A method for identifying genes which are differentially expressed in a normal healthy animal and an animal having a disease comprising:

a.   providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,
20   an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized;

25   b.   detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c.   detectably hybridizing to a second copy of said surface
30   polynucleotide sequences isolated from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d.   comparing the two hybridization patterns, wherein genes differentially expressed in a disease state are identified by the presence of differences
35   in the hybridization patterns at pre-defined regions;

e.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

5

7.   A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a.   providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy, uninfected animal and present in excess relative to the polynucleotide to be hybridized;

b.   providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from at least one selected cell, tissue, organ or organism sample of an infected animal;

c.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form first and second hybridization patterns on each said first and second surface,

d.   detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form third and fourth hybridization patterns on each said first and second surface,

e.   comparing the four hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

f.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

8.   A method for identifying a gene of a pathogen which is expressed in a biological sample of an animal infected with said pathogen comprising:

a.   providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample in of a healthy animal and an analogous selected sample of an animal having said disease and both present in excess relative to the polynucleotide to be hybridized

b.   detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c.   detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a sample from an infected animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d.   comparing the two hybridization patterns, wherein genes of said pathogen which are expressed in an infected animal are identified by the presence of differences in the hybridization patterns at pre-defined regions;

e.   identifying the oligonucleotide/polynucleotides on each surface which correspond to said pattern differences and the corresponding ESTs from which the oligonucleotide/polynucleotides are obtained, whereby identification of the EST permits identification of the gene from which the ESTs were derived.

9.   A composition suitable for use in hybridization comprising a solid surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences for hybridization, each sequence selected from the group consisting of a fragment of an EST, an entire EST a fragment of a gene or an entire gene isolated from a DNA library prepared from the group selected from at least one selected cell, tissue, organ or organism sample of a healthy animal, at least one analogous sample of said animal having a disease, at least one analogous sample of said animal infected with a microbial pathogen, and any combination thereof.

10.   An isolated gene sequence which is differentially expressed in a normal healthy animal and an animal having a disease, identified by the method of claim 1.

5   11.   An isolated pathogen gene sequence which is expressed in tissue or cell samples of an infected animal identified by the method of claim 7.

12.   A diagnostic composition useful for the diagnosis of a disease comprising a reagent capable of detectably targeting a gene sequence of claim 10 in a

10   biological sample of an animal.

13.   A diagnostic composition useful for the diagnosis of infection by a pathogen comprising a reagent capable of detectably targeting a gene sequence of claim 11 in a biological sample of an animal.

15

14.   An isolated protein produced by expression of a gene sequence of claim 10.

15.   An isolated pathogen protein produced by expression of a gene

20   sequence of claim 11.

16.   A therapeutic composition comprising a protein or fragment thereof selected from the group consisting of a protein of claim 10 and a protein of claim 15.

25   17.   A method for diagnosing a selected disease or infection in an animal comprising:
        a.   providing a first surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence selected from the group consisting of a fragment of an EST,

30   an entire EST a fragment of a gene or an entire gene, isolated from a DNA library prepared from at least one selected cell, tissue, organ or organism sample of a healthy animal and present in excess relative to the polynucleotide to be hybridized;
        b.   providing a second surface on which is immobilized at pre-defined regions of said surface a plurality of defined oligonucleotide/polynucleotide

35   sequences, each sequence comprising a fragment of an EST isolated from at least one said tissue of an animal having said disease;

27

c. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first and second hybridization pattern on each said first and second surface;

d. detectably hybridizing to a set of said first and second surfaces polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (c), said hybridization sufficient to form a third and fourth hybridization pattern on each said first and second surface;

e. comparing the four hybridization patterns, wherein substantial differences between the first and third hybridization patterns and the second and fourth hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and third hybridization patterns and second and fourth hybridization patterns indicates the absence of disease or infection.

18. A method for diagnosing a selected disease or infection in an animal comprising:

a. providing a surface on which is immobilized at pre-defined regions on said surface a plurality of defined oligonucleotide/polynucleotide sequences, each sequence comprising a fragment of an EST isolated from a DNA library prepared from the group consisting of a selected cell or tissue sample of a healthy animal and an analogous selected cell or tissue sample of an animal having said disease;

b. detectably hybridizing to a first copy of said surface polynucleotide sequences isolated from a sample from a healthy animal, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a first hybridization pattern on said surface;

c. detectably hybridizing to a second copy of said surface polynucleotide sequences isolated from a DNA library prepared from a sample from an animal having said disease, said sample selected from a cell or tissue sample analogous to the sample of step (a), said hybridization sufficient to form a second hybridization pattern on said surface;

d. comparing the two hybridization patterns, wherein substantial differences between the first and second hybridization patterns indicates the presence of said selected disease or infection in said animal, and substantial similarities in said first and second hybridization patterns indicates the absence of disease or infection.

28

INTERNATIONAL SEARCH REPORT

A.	CLASSIFICATION OF SUBJECT MATTER

IPC(6)   :C12Q 1/68
US CL   :435/6

According to International Patent Classification (IPC) or to both national classification and IPC

B.	FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. :   435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, CAS, BIOSIS

C.	DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	ANALYTICAL BIOCHEMISTRY, VOLUME 187, ISSUED 1990, FARGNOLI ET AL, "LOW-RATIO HYBRIDIZATION SUBTRACTION", PAGES 364-373, SEE ENTIRE DOCUMENT.	1-18
Y	PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES USA, VOLUME 88, ISSUED MARCH 1991, PATANJALI ET AL, "CONSTRUCTION OF A UNIFORM-ABUNDANCE (NORMALIZED) CDNA LIBRARY", PAGES 1943-1947, SEE ENTIRE DOCUMENT.	1-18
Y	SCIENCE, VOLUME 245, ISSUED 29 SEPTEMBER 1989, OLSON ET AL. "A COMMON LANGUAGE FOR PHYSICAL MAPPING OF THE HUMAN GENOME", PAGES 1434-1435, SEE ENTIRE DOCUMENT.	1-18

X	Further documents are listed in the continuation of Box C.		See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
03 APRIL 1995	17 MAY 1995

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	EGGERTON CAMPBELL
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	SCIENCE, VOLUME 252, ISSUED 21 JUNE 1991, ADAMS ET AL, "COMPLEMENTARY DNA SEQUENCING: EXPRESSED SEQUENCE TAGS AND HUMAN GENOME PROJECT", PAGES 1651-1656, SEE ENTIRE DOCUMENT.	1-18

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : C12Q 1/68, G06F 15/00	**A1**	**(11) International Publication Number:** WO 95/20681 **(43) International Publication Date:** 3 August 1995 (03.08.95)

(21) International Application Number: PCT/US95/01160

(22) International Filing Date: 27 January 1995 (27.01.95)

(30) Priority Data:
08/187,530   27 January 1994 (27.01.94)   US
08/282,955   29 July 1994 (29.07.94)   US

(71) Applicant: INCYTE PHARMACEUTICALS, INC. [US/US]; 3330 Hillview Avenue, Palo Alto, CA 94304 (US).

(72) Inventors: SEILHAMER, Jeffrey, J.; 12555 La Cresta, Los Altos Hills, CA 94022 (US). SCOTT, Randal, W.; 13140 Sun-Mor, Mountain View, CA 94040 (US).

(74) Agents: CAGE, Kenneth, L. et al.; Willian Brinks Hofer Gilson & Lione, 2000 K Street, N.W., Suite 200, Washington, DC 20006-1809 (US).

(81) Designated States: AM, AU, BB, BG, BR, BY, CA, CN, CZ, EE, FI, GE, HU, JP, KG, KP, KR, KZ, LK, LR, LT, LV, MD, MG, MN, MX, NO, NZ, PL, RO, RU, SI, SK, TJ, TT, UA, UZ, VN, European patent (AT, BE, CH, DE, DK, ES, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE), OAPI patent (BF, BJ, CF, CG, CI, CM, GA, GN, ML, MR, NE, SN, TD, TG), ARIPO patent (KE, MW, SD, SZ).

Published
With international search report.

(54) Title: COMPARATIVE GENE TRANSCRIPT ANALYSIS

(57) Abstract

A method and system for quantifying the relative abundance of gene transcripts in a biological specimen. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs (gene transcript imaging analysis). Another embodiment of the method produces a gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, the gene transcript imaging can be used to detect or diagnose a particular biological state, disease, or condition which is correlated to the relative abundance of gene transcripts in a given cell or population of cells. The invention provides a method for comparing the gene transcript image analysis from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes which are differentially expressed between the two specimens.

COMPARATIVE GENE TRANSCRIPT ANALYSIS

1. FIELD OF INVENTION

The present invention is in the field of molecular biology and computer science; more particularly, the
5 present invention describes methods of analyzing gene transcripts and diagnosing the genetic expression of cells and tissue.

2. BACKGROUND OF THE INVENTION

Until very recently, the history of molecular biology
10 has been written one gene at a time. Scientists have observed the cell's physical changes, isolated mixtures from the cell or its milieu, purified proteins, sequenced proteins and therefrom constructed probes to look for the corresponding gene.
15   Recently, different nations have set up massive projects to sequence the billions of bases in the human genome. These projects typically begin with dividing the genome into large portions of chromosomes and then determining the sequences of these pieces, which are then
20 analyzed for identity with known proteins or portions thereof, known as motifs. Unfortunately, the majority of genomic DNA does not encode proteins and though it is postulated to have some effect on the cell's ability to make protein, its relevance to medical applications is not
25 understood at this time.

A third methodology involves sequencing only the transcripts encoding the cellular machinery actively involved in making protein, namely the mRNA. The advantage is that the cell has already edited out all the non-coding
30 DNA, and it is relatively easy to identify the protein-coding portion of the RNA. The utility of this approach was not immediately obvious to genomic researchers. In fact, when cDNA sequencing was initially proposed, the method was roundly denounced by those committed to genomic
35 sequencing. For example, the head of the U.S. Human Genome project discounted cDNA sequencing as not valuable and refused to approve funding of projects.

In this disclosure, we teach methods for analyzing DNA, including cDNA libraries. Based on our analyses and

research, we see each individual gene product as a "pixel" of information, which relates to the expression of that, and only that, gene. We teach herein, methods whereby the individual "pixels" of gene expression information can be 5 combined into a single gene transcript "image," in which each of the individual genes can be visualized simultaneously and allowing relationships between the gene pixels to be easily visualized and understood.

We further teach a new method which we call electronic 10 subtraction. Electronic subtraction will enable the gene researcher to turn a single image into a moving picture, one which describes the temporality or dynamics of gene expression, at the level of a cell or a whole tissue. It is that sense of "motion" of cellular machinery on the 15 scale of a cell or organ which constitutes the new invention herein. This constitutes a new view into the process of living cell physiology and one which holds great promise to unveil and discover new therapeutic and diagnostic approaches in medicine.

20 We teach another method which we call "electronic northern," which tracks the expression of a single gene across many types of cells and tissues.

Nucleic acids (DNA and RNA) carry within their sequence the hereditary information and are therefore the 25 prime molecules of life. Nucleic acids are found in all living organisms including bacteria, fungi, viruses, plants and animals. It is of interest to determine the relative abundance of different discrete nucleic acids in different cells, tissues and organisms over time under various 30 conditions, treatments and regimes.

All dividing cells in the human body contain the same set of 23 pairs of chromosomes. It is estimated that these autosomal and sex chromosomes encode approximately 100,000 genes. The differences among different types of cells are 35 believed to reflect the differential expression of the 100,000 or so genes. Fundamental questions of biology could be answered by understanding which genes are transcribed and knowing the relative abundance of transcripts in different cells.

Previously, the art has only provided for the analysis
of a few known genes at a time by standard molecular
biology techniques such as PCR, northern blot analysis, or
other types of DNA probe analysis such as _in situ_
5 hybridization. Each of these methods allows one to analyze
the transcription of only known genes and/or small numbers
of genes at a time. Nucl. Acids Res. 19, 7097-7104 (1991);
Nucl. Acids Res. 18, 4833-42 (1990); Nucl. Acids Res. 18,
2789-92 (1989); European J. Neuroscience 2, 1063-1073
10 (1990); Analytical Biochem. 187, 364-73 (1990); Genet.
Annals Techn. Appl. 7, 64-70 (1990); GATA 8(4), 129-33
(1991); Proc. Natl. Acad. Sci. USA 85, 1696-1700 (1988);
Nucl. Acids Res. 19, 1954 (1991); Proc. Natl. Acad. Sci.
USA 88, 1943-47 (1991); Nucl. Acids Res. 19, 6123-27
15 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42 (1988);
Nucl. Acids Res. 16, 10937 (1988).

Studies of the number and types of genes whose
transcription is induced or otherwise regulated during cell
processes such as activation, differentiation, aging, viral
20 transformation, morphogenesis, and mitosis have been
pursued for many years, using a variety of methodologies.
One of the earliest methods was to isolate and analyze
levels of the proteins in a cell, tissue, organ system, or
even organisms both before and after the process of
25 interest. One method of analyzing multiple proteins in a
sample is using 2-dimensional gel electrophoresis, wherein
proteins can be, in principle, identified and quantified as
individual bands, and ultimately reduced to a discrete
signal. At present, 2-dimensional analysis only resolves
30 approximately 15% of the proteins. In order to positively
analyze those bands which are resolved, each band must be
excised from the membrane and subjected to protein sequence
analysis using Edman degradation. Unfortunately, most of
the bands were present in quantities too small to obtain a
35 reliable sequence, and many of those bands contained more
than one discrete protein. An additional difficulty is
that many of the proteins were blocked at the
amino-terminus, further complicating the sequencing
process.

Analyzing differentiation at the gene transcription
level has overcome many of these disadvantages and
drawbacks, since the power of recombinant DNA technology
allows amplification of signals containing very small
5   amounts of material. The most common method, called
"hybridization subtraction," involves isolation of mRNA
from the biological specimen before (B) and after (A) the
developmental process of interest, transcribing one set of
mRNA into cDNA, subtracting specimen B from specimen A
10   (mRNA from cDNA) by hybridization, and constructing a cDNA
library from the non-hybridizing mRNA fraction. Many
different groups have used this strategy successfully, and
a variety of procedures have been published and improved
upon using this same basic scheme. Nucl. Acids Res. 19,
15   7097-7104 (1991); Nucl. Acids Res. 18, 4833-42 (1990);
Nucl. Acids Res. 18, 2789-92 (1989); European J.
Neuroscience 2, 1063-1073 (1990); Analytical Biochem. 187,
364-73 (1990); Genet. Annals Techn. Appl. 7, 64-70 (1990);
GATA 8(4), 129-33 (1991); Proc. Natl. Acad. Sci. USA 85,
20   1696-1700 (1988); Nucl. Acids Res. 19, 1954 (1991); Proc.
Natl. Acad. Sci. USA 88, 1943-47 (1991); Nucl. Acids Res.
19, 6123-27 (1991); Proc. Natl. Acad. Sci. USA 85, 5738-42
(1988); Nucl. Acids Res. 16, 10937 (1988).

Although each of these techniques have particular
25   strengths and weaknesses, there are still some limitations
and undesirable aspects of these methods: First, the time
and effort required to construct such libraries is quite
large. Typically, a trained molecular biologist might
expect construction and characterization of such a library
30   to require 3 to 6 months, depending on the level of skill,
experience, and luck. Second, the resulting subtraction
libraries are typically inferior to the libraries
constructed by standard methodology. A typical
conventional cDNA library should have a clone complexity of
35   at least 10^6 clones, and an average insert size of 1-3 kB.
In contrast, subtracted libraries can have complexities of
10^2 or 10^3 and average insert sizes of 0.2 kB. Therefore,
there can be a significant loss of clone and sequence
information associated with such libraries. Third, this

approach allows the researcher to capture only the genes
induced in specimen A relative to specimen B, not
vice-versa, nor does it easily allow comparison to a third
specimen of interest (C). Fourth, this approach requires
5 very large amounts (hundreds of micrograms) of "driver"
mRNA (specimen B), which significantly limits the number
and type of subtractions that are possible since many
tissues and cells are very difficult to obtain in large
quantities.

10   Fifth, the resolution of the subtraction is dependent
upon the physical properties of DNA:DNA or RNA:DNA
hybridization. The ability of a given sequence to find a
hybridization match is dependent on its unique CoT value.
The CoT value is a function of the number of copies
15 (concentration) of the particular sequence, multiplied by
the time of hybridization. It follows that for sequences
which are abundant, hybridization events will occur very
rapidly (low CoT value), while rare sequences will form
duplexes at very high CoT values. CoT values which allow
20 such rare sequences to form duplexes and therefore be
effectively selected are difficult to achieve in a
convenient time frame. Therefore, hybridization
subtraction is simply not a useful technique with which to
study relative levels of rare mRNA species. Sixth, this
25 problem is further complicated by the fact that duplex
formation is also dependent on the nucleotide base
composition for a given sequence. Those sequences rich in
G + C form stronger duplexes than those with high contents
of A + T. Therefore, the former sequences will tend to be
30 removed selectively by hybridization subtraction. Seventh,
it is possible that hybridization between nonexact matches
can occur. When this happens, the expression of a
homologous gene may "mask" expression of a gene of
interest, artificially skewing the results for that
35 particular gene.

       Matsubara and Okubo proposed using partial cDNA
sequences to establish expression profiles of genes which
could be used in functional analyses of the human genome.
Matsubara and Okubo warned against using random priming, as

it creates multiple unique DNA fragments from individual
mRNAs and may thus skew the analysis of the number of
particular mRNAs per library. They sequenced randomly
selected members from a 3'-directed cDNA library and

5 established the frequency of appearance of the various
ESTs. They proposed comparing lists of ESTs from various
cell types to classify genes. Genes expressed in many
different cell types were labeled housekeepers and those
selectively expressed in certain cells were labeled cell-

10 specific genes, even in the absence of the full sequence of
the gene or the biological activity of the gene product.

The present invention avoids the drawbacks of the
prior art by providing a method to quantify the relative
abundance of multiple gene transcripts in a given

15 biological specimen by the use of high-throughput
sequence-specific analysis of individual RNAs and/or their
corresponding cDNAs.

The present invention offers several advantages over
current protein discovery methods which attempt to isolate

20 individual proteins based upon biological effects. The
method of the instant invention provides for detailed
diagnostic comparisons of cell profiles revealing numerous
changes in the expression of individual transcripts.

The instant invention provides several advantages over

25 current subtraction methods including a more complex
library analysis (10^6 to 10^7 clones as compared to 10^3
clones) which allows identification of low abundance
messages as well as enabling the identification of messages
which either increase or decrease in abundance. These

30 large libraries are very routine to make in contrast to the
libraries of previous methods. In addition, homologues can
easily be distinguished with the method of the instant
invention.

This method is very convenient because it organizes a

35 large quantity of data into a comprehensible, digestible
format. The most significant differences are highlighted
by electronic subtraction. In depth analyses are made more
convenient.

The pres nt invention provides several advantages over previous methods of electronic analysis of cDNA. The method is particularly powerful when more than 100 and preferably more than 1,000 gene transcripts are analyzed.

5   In such a case, new low-frequency transcripts are discovered and tissue typed.

High resolution analysis of gene expression can be used directly as a diagnostic profile or to identify disease-specific genes for the development of more classic

10   diagnostic approaches.

This process is defined as gene transcript frequency analysis. The resulting quantitative analysis of the gene transcripts is defined as comparative gene transcript analysis.

15   3.   **SUMMARY OF THE INVENTION**

The invention is a method of analyzing a specimen containing gene transcripts comprising the steps of (a) producing a library of biological sequences; (b) generating a set of transcript sequences, where each of the transcript

20   sequences in said set is indicative of a different one of the biological sequences of the library; (c) processing the transcript sequences in a programmed computer (in which a database of reference transcript sequences indicative of reference sequences is stored), to generate an identified

25   sequence value for each of the transcript sequences, where each said identified sequence value is indicative of sequence annotation and a degree of match between one of the biological sequences of the library and at least one of the reference sequences; and (d) processing each said

30   identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

The invention also includes a method of comparing two specimens containing gene transcripts. The first specimen

35   is processed as described above. The second specimen is used to produce a second library of biological sequences, which is used to generate a second set of transcript sequences, where each of the transcript sequences in the

second set is indicative of one of the biological sequences of the second library. Then the second set of transcript sequences is processed in a programmed computer to generate a second set of identified sequence values, namely the

5   further identified sequence values, each of which is indicative of a sequence annotation and includes a degree of match between one of the biological sequences of the second library and at least one of the reference sequences. The further identified sequence values are processed to

10 generate further final data values indicative of the number of times each further identified sequence value is present in the second library. The final data values from the first specimen and the further identified sequence values from the second specimen are processed to generate ratios

15 of transcript sequences, which indicate the differences in the number of gene transcripts between the two specimens.

In a further embodiment, the method includes quantifying the relative abundance of mRNA in a biological specimen by (a) isolating a population of mRNA transcripts

20 from a biological specimen; (b) identifying genes from which the mRNA was transcribed by a sequence-specific method; (c) determining the numbers of mRNA transcripts corresponding to each of the genes; and (d) using the mRNA transcript numbers to determine the relative abundance of

25 mRNA transcripts within the population of mRNA transcripts.

Also disclosed is a method of producing a gene transcript image analysis by first obtaining a mixture of mRNA, from which cDNA copies are made. The cDNA is inserted into a suitable vector which is used to transfect

30 suitable host strain cells which are plated out and permitted to grow into clones, each cone representing a unique mRNA. A representative population of clones transfected with cDNA is isolated. Each clone in the population is identified by a sequence-specific method

35 which identifies the gene from which the unique mRNA was transcribed. The number of times each gene is identified to a clone is determined to evaluate gene transcript abundance. The genes and their abundances are listed in order of abundance to produce a gene transcript image.

In a further embodiment, the relative abundance of the gene transcripts in one cell type or tissue is compared with the relative abundance of gene transcript numbers in a second cell type or tissue in order to identify the

5   differences and similarities.

In a further embodiment, the method includes a system for analyzing a library of biological sequences including a means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a

10   different one of the biological sequences of the library; and a means for processing the transcript sequences in a computer system in which a database of reference transcript sequences indicative of reference sequences is stored, wherein the computer is programmed with software for

15   generating an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and the degree of match between a different one of the biological sequences of the library and at least one of the reference

20   sequences, and for processing each said identified sequence value to generate final data values indicative of the number of times each identified sequence value is present in the library.

In essence, the invention is a method and system for

25   quantifying the relative abundance of gene transcripts in a biological specimen. The invention provides a method for comparing the gene transcript image from two or more different biological specimens in order to distinguish between the two specimens and identify one or more genes

30   which are differentially expressed between the two specimens. Thus, this gene transcript image and its comparison can be used as a diagnostic. One embodiment of the method generates high-throughput sequence-specific analysis of multiple RNAs or their corresponding cDNAs: a

35   gene transcript image. Another embodiment of the method produces the gene transcript imaging analysis by the use of high-throughput cDNA sequence analysis. In addition, two or more gene transcript images can be compared and used to detect or diagnose a particular biological state, disease,

9

or condition which is correlated to the relative abundance
of gene transcripts in a given cell or population of cells.

4. DESCRIPTION OF THE TABLES AND DRAWINGS
4.1. TABLES

5   Table 1 presents a detailed explanation of the letter
codes utilized in Tables 2-5.

Table 2 lists the one hundred most common gene
transcripts. It is a partial list of isolates from the
HUVEC cDNA library prepared and sequenced as described

10   below. The left-hand column refers to the sequence's order
of abundance in this table. The next column labeled
"number" is the clone number of the first HUVEC sequence
identification reference matching the sequence in the
"entry" column number. Isolates that have not been

15   sequenced are not present in Table 2. The next column,
labeled "N", indicates the total number of cDNAs which have
the same degree of match with the sequence of the reference
transcript in the "entry" column.

The column labeled "entry" gives the NIH GENBANK locus

20   name, which corresponds to the library sequence numbers.
The "s" column indicates in a few cases the species of the
reference sequence. The code for column "s" is given in
Table 1. The column labeled "descriptor" provides a plain
English explanation of the identity of the sequence

25   corresponding to the NIH GENBANK locus name in the "entry"
column.

Table 3 is a comparison of the top fifteen most
abundant gene transcripts in normal monocytes and activated
macrophage cells.

30   Table 4 is a detailed summary of library subtraction
analysis summary comparing the THP-1 and human macrophage
cDNA sequences. In Table 4, the same code as in Table 2 is
used. Additional columns are for "bgfreq" (abundance
number in the subtractant library), "rfend" (abundance

35   number in the target library) and "ratio" (the target
abundance number divided by the subtractant abundance
number). As is clear from perusal of the table, when the
abundance number in the subtractant library is "0", the

target abundance number is divided by 0.05. This is a way
of obtaining a result (not possible dividing by 0) and
distinguishing the result from ratios of subtractant
numbers of 1.

5   Table 5 is the computer program, written in source
code, for generating gene transcript subtraction profiles.

       Table 6 is a partial listing of database entries used
in the electronic northern blot analysis as provided by the
present invention.

10

4.2. BRIEF DESCRIPTION OF THE DRAWINGS

       Figure 1 is a chart summarizing data collected and
stored regarding the library construction portion of
sequence preparation and analysis.

15   Figure 2 is a diagram representing the sequence of
operations performed by "abundance sort" software in a
class of preferred embodiments of the inventive method.

       Figure 3 is a block diagram of a preferred embodiment
of the system of the invention.

20   Figure 4 is a more detailed block diagram of the
bioinformatics process from new sequence (that has already
been sequenced but not identified) to printout of the
transcript imaging analysis and the provision of database
subscriptions.

25   5. DETAILED DESCRIPTION OF THE INVENTION

       The present invention provides a method to compare the
relative abundance of gene transcripts in different
biological specimens by the use of high-throughput
sequence-specific analysis of individual RNAs or their
30 corresponding cDNAs (or alternatively, of data representing
other biological sequences). This process is denoted
herein as gene transcript imaging. The quantitative
analysis of the relative abundance for a set of gene
transcripts is denoted herein as "gene transcript image
35 analysis" or "gene transcript frequency analysis". The
present invention allows one to obtain a profile for gene
transcription in any given population of cells or tissue
from any type of organism. The invention can be applied to

obtain a profile of a specimen consisting of a single cell
(or clones of a single cell), or of many cells, or of
tissue more complex than a single cell and containing
multiple cell types, such as liver.

5   The invention has significant advantages in the fields
of diagnostics, toxicology and pharmacology, to name a few.
A highly sophisticated diagnostic test can be performed on
the ill patient in whom a diagnosis has not been made. A
biological specimen consisting of the patient's fluids or

10   tissues is obtained, and the gene transcripts are isolated
and expanded to the extent necessary to determine their
identity. Optionally, the gene transcripts can be
converted to cDNA. A sampling of the gene transcripts are
subjected to sequence-specific analysis and quantified.

15   These gene transcript sequence abundances are compared
against reference database sequence abundances including
normal data sets for diseased and healthy patients. The
patient has the disease(s) with which the patient's data
set most closely correlates.

20   For example, gene transcript frequency analysis can be
used to differentiate normal cells or tissues from diseased
cells or tissues, just as it highlights differences between
normal monocytes and activated macrophages in Table 3.

           In toxicology, a fundamental question is which tests

25   are most effective in predicting or detecting a toxic
effect. Gene transcript imaging provides highly detailed
information on the cell and tissue environment, some of
which would not be obvious in conventional, less detailed
screening methods. The gene transcript image is a more

30   powerful method to predict drug toxicity and efficacy.
Similar benefits accrue in the use of this tool in
pharmacology. The gene transcript image can be used
selectively to look at protein categories which are
expected to be affected, for example, enzymes which

35   detoxify toxins.

           In an alternative embodiment, comparative gene
transcript frequency analysis is used to differentiate
between cancer cells which respond to anti-cancer agents
and those which do not respond. Examples of anti-cancer

agents are tamoxifen, vincristine, vinblastine, podophyllotoxins, etoposide, tenisposide, cisplatin, biologic response modifiers such as interferon, Il-2, GM-CSF, enzymes, hormones and the like. This method also

5 provides a means for sorting the gene transcripts by functional category. In the case of cancer cells, transcription factors or other essential regulatory molecules are very important categories to analyze across different libraries.

10   In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between control liver cells and liver cells isolated from patients treated with experimental drugs like FIAU to distinguish between pathology caused by the underlying disease and that caused

15 by the drug.

    In yet another embodiment, comparative gene transcript frequency analysis is used to differentiate between brain tissue from patients treated and untreated with lithium.

    In a further embodiment, comparative gene transcript

20 frequency analysis is used to differentiate between cyclosporin and FK506-treated cells and normal cells.

    In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between virally infected (including HIV-infected) human cells and

25 uninfected human cells. Gene transcript frequency analysis is also used to rapidly survey gene transcripts in HIV-resistant, HIV-infected, and HIV-sensitive cells. Comparison of gene transcript abundance will indicate the success of treatment and/or new avenues to study.

30   In a further embodiment, comparative gene transcript frequency analysis is used to differentiate between bronchial lavage fluids from healthy and unhealthy patients with a variety of ailments.

    In a further embodiment, comparative gene transcript

35 frequency analysis is used to differentiate between cell, plant, microbial and animal mutants and wild-type species. In addition, the transcript abundance program is adapted to permit the scientist to evaluate the transcription of one gene in many different tissues. Such comparisons could

identify deletion mutants which do not produce a gene
product and point mutants which produce a less abundant or
otherwise different message. Such mutations can affect
basic biochemical and pharmacological processes, such as
5 mineral nutrition and metabolism, and can be isolated by
means known to those skilled in the art. Thus, crops with
improved yields, pest resistance and other factors can be
developed.

In a further embodiment, comparative gene transcript
10 frequency analysis is used for an interspecies comparative
analysis which would allow for the selection of better
pharmacologic animal models. In this embodiment, humans
and other animals (such as a mouse), or their cultured
cells are treated with a specific test agent. The relative
15 sequence abundance of each cDNA population is determined.
If the animal test system is a good model, homologous genes
in the animal cDNA population should change expression
similarly to those in human cells. If side effects are
detected with the drug, a detailed transcript abundance
20 analysis will be performed to survey gene transcript
changes. Models will then be evaluated by comparing basic
physiological changes.

In a further embodiment, comparative gene transcript
frequency analysis is used in a clinical setting to give a
25 highly detailed gene transcript profile of a patient's
cells or tissue (for example, a blood sample). In
particular, gene transcript frequency analysis is used to
give a high resolution gene expression profile of a
diseased state or condition.

30 In the preferred embodiment, the method utilizes
high-throughput cDNA sequencing to identify specific
transcripts of interest. The generated cDNA and deduced
amino acid sequences are then extensively compared with
GENBANK and other sequence data banks as described below.
35 The method offers several advantages over current protein
discovery by two-dimensional gel methods which try to
identify individual proteins involved in a particular
biological effect. Here, detailed comparisons of profiles
of activated and inactive cells reveal numerous changes in

the expression of individual transcripts. After it is
determined if the sequence is an "exact" match, similar or
a non-match, the sequence is ntered into a database.
Next, the numbers of copies of cDNA corresponding to each
5   gene are tabulated. Although this can be done slowly and
arduously, if at all, by human hand from a printout of all
entries, a computer program is a useful and rapid way to
tabulate this information. The numbers of cDNA copies
(optionally divided by the total number of sequences in the
10 data set) provides a picture of the relative abundance of
transcripts for each corresponding gene. The list of
represented genes can then be sorted by abundance in the
cDNA population. A multitude of additional types of
comparisons or dimensions are possible and are exemplified
15 below.

An alternate method of producing a gene transcript
image includes the steps of obtaining a mixture of test
mRNA and providing a representative array of unique probes
whose sequences are complementary to at least some of the
20 test mRNAs. Next, a fixed amount of the test mRNA is added
to the arrayed probes. The test mRNA is incubated with the
probes for a sufficient time to allow hybrids of the test
mRNA and probes to form. The mRNA-probe hybrids are
detected and the quantity determined. The hybrids are
25 identified by their location in the probe array. The
quantity of each hybrid is summed to give a population
number. Each hybrid quantity is divided by the population
number to provide a set of relative abundance data termed a
gene transcript image analysis.

30   6. EXAMPLES
The examples below are provided to illustrate the
subject invention. These examples are provided by way of
illustration and are not included for the purpose of
limiting the invention.

35   6.1. TISSUE SOURCES AND CELL LINES
For analysis with the computer program claimed herein,
biological sequences can be obtained from virtually any

15

source. Most popular are tissues obtained from the human body. Tissues can be obtained from any organ of the body, any age donor, any abnormality or any immortalized cell lin . Immortal cell lines may be preferred in some

5   instances because of their purity of cell type; other tissue samples invariably include mixed cell types. A special technique is available to take a single cell (for example, a brain cell) and harness the cellular machinery to grow up sufficient cDNA for sequencing by the techniques

10   and analysis described herein (cf. U.S. Patent Nos. 5,021,335 and 5,168,038, which are incorporated by reference). The examples given herein utilized the following immortalized cell lines: monocyte-like U-937 cells, activated macrophage-like THP-1 cells, induced

15   vascular endothelial cells (HUVEC cells) and mast cell-like HMC-1 cells.

The U-937 cell line is a human histiocytic lymphoma cell line with monocyte characteristics, established from malignant cells obtained from the pleural effusion of a

20   patient with diffuse histiocytic lymphoma (Sundstrom, C. and Nilsson, K. (1976) Int. J. Cancer 17:565). U-937 is one of only a few human cell lines with the morphology, cytochemistry, surface receptors and monocyte-like characteristics of histiocytic cells. These cells can be

25   induced to terminal monocytic differentiation and will express new cell surface molecules when activated with supernatants from human mixed lymphocyte cultures. Upon this type of _in vitro_ activation, the cells undergo morphological and functional changes, including

30   augmentation of antibody-dependent cellular cytotoxicity (ADCC) against erythroid and tumor target cells (one of the principal functions of macrophages). Activation of U-937 cells with phorbol 12-myristate 13-acetate (PMA) _in vitro_ stimulates the production of several compounds, including

35   prostaglandins, leukotrienes and platelet-activating factor (PAF), which are potent inflammatory mediators. Thus, U-937 is a cell line that is well suited for the identification and isolation of gene transcripts associated with normal monocytes.

The HUVEC cell line is a normal, homogeneous, well characterized, early passage endothelial cell culture from human umbilical vein (Cell Systems Corp., 12815 NE 124th Street, Kirkland, WA 98034). Only gene transcripts from
5 induced, or treated, HUVEC cells were sequenced. One batch of 1 X 10^8 cells was treated for 5 hours with 1 U/ml rIL-1b and 100 ng/ml E.coli lipopolysaccharide (LPS) endotoxin prior to harvesting. A separate batch of 2 X 10^8 cells was treated at confluence with 4 U/ml TNF and 2 U/ml
10 interferon-gamma (IFN-gamma) prior to harvesting.

THP-1 is a human leukemic cell line with distinct monocytic characteristics. This cell line was derived from the blood of a 1-year-old boy with acute monocytic leukemia (Tsuchiya, S. et al. (1980) Int. J. Cancer: 171-76). The
15 following cytological and cytochemical criteria were used to determine the monocytic nature of the cell line: 1) the presence of alpha-naphthyl butyrate esterase activity which could be inhibited by sodium fluoride; 2) the production of lysozyme; 3) the phagocytosis of latex particles and
20 sensitized SRBC (sheep red blood cells); and 4) the ability of mitomycin C-treated THP-1 cells to activate T-lymphocytes following ConA (concanavalin A) treatment. Morphologically, the cytoplasm contained small azurophilic granules and the nucleus was indented and irregularly
25 shaped with deep folds. The cell line had Fc and C3b receptors, probably functioning in phagocytosis. THP-1 cells treated with the tumor promoter 12-o-tetradecanoyl-phorbol-13 acetate (TPA) stop proliferating and differentiate into macrophage-like cells which mimic native
30 monocyte-derived macrophages in several respects. Morphologically, as the cells change shape, the nucleus becomes more irregular and additional phagocytic vacuoles appear in the cytoplasm. The differentiated THP-1 cells also exhibit an increased adherence to tissue culture
35 plastic.

HMC-1 cells (a human mast cell line) were established from the peripheral blood of a Mayo Clinic patient with mast cell leukemia (Leukemia Res. (1988) 12:345-55). The cultured cells looked similar to immature cloned murine

17

mast cells, contained histamine, and stained positively for
chloroacetate esterase, amino caproate esterase, eosinophil
major basic protein (MBP) and tryptase. The HMC-1 cells
have, however, lost the ability to synthesize normal IgE

5   receptors. HMC-1 cells also possess a 10;16 translocation,
present in cells initially collected by leukophoresis from
the patient and not an artifact of culturing. Thus, HMC-1
cells are a good model for mast cells.

6.2. CONSTRUCTION OF cDNA LIBRARIES

10   For inter-library comparisons, the libraries must be
prepared in similar manners. Certain parameters appear to
be particularly important to control. One such parameter
is the method of isolating mRNA. It is important to use
the same conditions to remove DNA and heterogeneous nuclear

15 RNA from comparison libraries. Size fractionation of cDNA
must be carefully controlled. The same vector preferably
should be used for preparing libraries to be compared. At
the very least, the same type of vector (e.g.,
unidirectional vector) should be used to assure a valid

20 comparison. A unidirectional vector may be preferred in
order to more easily analyze the output.
        It is preferred to prime only with oligo dT
unidirectional primer in order to obtain one only clone per
mRNA transcript when obtaining cDNAs. However, it is

25 recognized that employing a mixture of oligo dT and random
primers can also be advantageous because such a mixture
results in more sequence diversity when gene discovery also
is a goal. Similar effects can be obtained with DR2
(Clontech) and HXLOX (US Biochemical) and also vectors from

30 Invitrogen and Novagen. These vectors have two
requirements. First, there must be primer sites for
commercially available primers such as T3 or M13 reverse
primers. Second, the vector must accept inserts up to 10
kB.

35   It also is important that the clones be randomly
sampled, and that a significant population of clones is
used. Data have been generated with 5,000 clones; however,
if very rare genes are to be obtained and/or their relative

abundance determined, as many as 100,000 clones from a single library may need to be sampled. Size fractionation of cDNA also must be carefully controlled. Alternately, plaques can be selected, rather than clones.

5   Besides the Uni-ZAP™ vector system by Stratagene disclosed below, it is now believed that other similarly unidirectional vectors also can be used. For example, it is believed that such vectors include but are not limited to DR2 (Clontech), and HXLOX (U.S. Biochemical).

10   Preferably, the details of library construction (as shown in Figure 1) are collected and stored in a database for later retrieval relative to the sequences being compared. Fig. 1 shows important information regarding the library collaborator or cell or cDNA supplier,
15 pretreatment, biological source, culture, mRNA preparation and cDNA construction. Similarly detailed information about the other steps is beneficial in analyzing sequences and libraries in depth.

RNA must be harvested from cells and tissue samples
20 and cDNA libraries are subsequently constructed. cDNA libraries can be constructed according to techniques known in the art. (See, for example, Maniatis, T. et al. (1982) Molecular Cloning, Cold Spring Harbor Laboratory, New York). cDNA libraries may also be purchased. The U-937
25 cDNA library (catalog No. 937207) was obtained from Stratagene, Inc., 11099 M. Torrey Pines Rd., La Jolla, CA 92037.

The THP-1 cDNA library was custom constructed by Stratagene from THP-1 cells cultured 48 hours with 100 nm
30 TPA and 4 hours with 1 μg/ml LPS. The human mast cell HMC-1 cDNA library was also custom constructed by Stratagene from cultured HMC-1 cells. The HUVEC cDNA library was custom constructed by Stratagene from two batches of induced HUVEC cells which were separately processed.

35   Essentially, all the libraries were prepared in the same manner. First, poly(A+)RNA (mRNA) was purified. For the U-937 and HMC-1 RNA, cDNA synthesis was only primed with oligo dT. For the THP-1 and HUVEC RNA, cDNA synthesis was primed separately with both oligo dT and random

19

hexamers, and the two cDNA libraries were treated
separately. Synthetic adaptor oligonucleotides were
ligated onto cDNA ends enabling its insertion into the Uni-
Zap™ vector system (Stratagene), allowing high efficiency
5 unidirectional (sense orientation) lambda library
construction and the convenience of a plasmid system with
blue-white color selection to detect clones with cDNA
insertions. Finally, the two libraries were combined into
a single library by mixing equal numbers of bacteriophage.
10   The libraries can be screened with either DNA probes
or antibody probes and the pBluescript® phagemid
(Stratagene) can be rapidly excised _in vivo_. The phagemid
allows the use of a plasmid system for easy insert
characterization, sequencing, site-directed mutagenesis,
15 the creation of unidirectional deletions and expression of
fusion proteins. The custom-constructed library phage
particles were infected into _E. coli_ host strain XL1-Blue®
(Stratagene), which has a high transformation efficiency,
increasing the probability of obtaining rare, under-
20 represented clones in the cDNA library.

6.3.   ISOLATION OF cDNA CLONES

The phagemid forms of individual cDNA clones were
obtained by the _in vivo_ excision process, in which the host
bacterial strain was coinfected with both the lambda
25 library phage and an f1 helper phage. Proteins derived
from both the library-containing phage and the helper phage
nicked the lambda DNA, initiated new DNA synthesis from
defined sequences on the lambda target DNA and created a
smaller, single stranded circular phagemid DNA molecule
30 that included all DNA sequences of the pBluescript® plasmid
and the cDNA insert. The phagemid DNA was secreted from
the cells and purified, then used to re-infect fresh host
cells, where the double stranded phagemid DNA was produced.
Because the phagemid carries the gene for beta-lactamase,
35 the newly-transformed bacteria are selected on medium
containing ampicillin.

Phagemid DNA was purified using the Magic Minipreps™
DNA Purification System (Promega catalogue #A7100. Promega

20

Corp., 2800 Woods Hollow Rd., Madison, WI 53711). This
small-scale process provides a simple and reliable method
f r lysing the bacterial cells and rapidly isolating
purified phagemid DNA using a proprietary DNA-binding
5   resin. The DNA was eluted from the purification resin
already prepared for DNA sequencing and other analytical
manipulations.

Phagemid DNA was also purified using the QIAwell-8
Plasmid Purification System from QIAGEN® DNA Purification
10   System (QIAGEN Inc., 9259 Eton Ave., Chattsworth, CA
91311). This product line provides a convenient, rapid and
reliable high-throughput method for lysing the bacterial
cells and isolating highly purified phagemid DNA using
QIAGEN anion-exchange resin particles with EMPORE™ membrane
15   technology from 3M in a multiwell format. The DNA was
eluted from the purification resin already prepared for DNA
sequencing and other analytical manipulations.

An alternate method of purifying phagemid has recently
become available. It utilizes the Miniprep Kit (Catalog
20   No. 77468, available from Advanced Genetic Technologies
Corp., 19212 Orbit Drive, Gaithersburg, Maryland). This
kit is in the 96-well format and provides enough reagents
for 960 purifications. Each kit is provided with a
recommended protocol, which has been employed except for
25   the following changes. First, the 96 wells are each filled
with only 1 ml of sterile terrific broth with carbenicillin
at 25 mg/L and glycerol at 0.4%. After the wells are
inoculated, the bacteria are cultured for 24 hours and
lysed with 60 μl of lysis buffer. A centrifugation step
30   (2900 rpm for 5 minutes) is performed before the contents
of the block are added to the primary filter plate. The
optional step of adding isopropanol to TRIS buffer is not
routinely performed. After the last step in the protocol,
samples are transferred to a Beckman 96-well block for
35   storage.

Another new DNA purification system is the WIZARD™
product line which is available from Promega (catalog No.
A7071) and may be adaptable to the 96-well format.

6.4. SEQUENCING OF cDNA CLONES

The cDNA inserts from random isolates of the U-937 and
THP-1 libraries were sequenced in part. Methods for DNA
sequencing are well known in the art. Conventional

5 enzymatic methods employ DNA polymerase Klenow fragment,
Sequenase™ or Taq polymerase to extend DNA chains from an
oligonucleotide primer annealed to the DNA template of
interest. Methods have been developed for the use of both
single- and double-stranded templates. The chain

10 termination reaction products are usually electrophoresed
on urea-acrylamide gels and are detected either by
autoradiography (for radionuclide-labeled precursors) or by
fluorescence (for fluorescent-labeled precursors). Recent
improvements in mechanized reaction preparation, sequencing

15 and analysis using the fluorescent detection method have
permitted expansion in the number of sequences that can be
determined per day (such as the Applied Biosystems 373 and
377 DNA sequencer, Catalyst 800). Currently with the
system as described, read lengths range from 250 to 400

20 bases and are clone dependent. Read length also varies
with the length of time the gel is run. In general, the
shorter runs tend to truncate the sequence. A minimum of
only about 25 to 50 bases is necessary to establish the
identification and degree of homology of the sequence.

25 Gene transcript imaging can be used with any sequence-
specific method, including, but not limited to
hybridization, mass spectroscopy, capillary electrophoresis
and 505 gel electrophoresis.

6.5. HOMOLOGY SEARCHING OF cDNA CLONE AND
30   DEDUCED PROTEIN (and Subsequent Steps)

Using the nucleotide sequences derived from the cDNA
clones as query sequences (sequences of a Sequence
Listing), databases containing previously identified
sequences are searched for areas of homology (similarity).

35 Examples of such databases include Genbank and EMBL. We
next describe examples of two homology search algorithms
that can be used, and then describe the subsequent
computer-implemented steps to be performed in accordance
with preferred embodiments of the invention.

In the following description of the computer-
implemented steps of the invention, the word "library"
denotes a set (or population) of biological specimen
nucleic acid sequences. A "library" can consist of cDNA

5   sequences, RNA sequences, or the like, which characterize a
biological specimen. The biological specimen can consist
of cells of a single human cell type (or can be any of the
other above-mentioned types of specimens). We contemplate
that the sequences in a library have been determined so as

10 to accurately represent or characterize a biological
specimen (for example, they can consist of representative
cDNA sequences from clones of RNA taken from a single human
cell).

In the following description of the computer-

15 implemented steps of the invention, the expression
"database" denotes a set of stored data which represent a
collection of sequences, which in turn represent a
collection of biological reference materials. For example,
a database can consist of data representing many stored

20 cDNA sequences which are in turn representative of human
cells infected with various viruses, cells of humans of
various ages, cells from different mammalian species, and
so on.

In preferred embodiments, the invention employs a

25 computer programmed with software (to be described) for
performing the following steps:

(a) processing data indicative of a library of cDNA
sequences (generated as a result of high-throughput cDNA
sequencing or other method) to determine whether each

30 sequence in the library matches a DNA sequence of a
reference database of DNA sequences (and if so, identifying
the reference database entry which matches the sequence and
indicating the degree of match between the reference
sequence and the library sequence) and assigning an

35 identified sequence value based on the sequence annotation
and degree of match to each of the sequences in the
library;

(b) for some or all entries of the database,
tabulating the number of matching identified sequence

values in the library (Although this can be done by human
hand from a printout of all entries, we prefer to perform
this step using computer software to be described below.),
thereby generating a set of final data values or "abundance

5   numbers"; and

        (c) if the libraries are different sizes, dividing
each abundance number by the total number of sequences in
the library, to obtain a relative abundance number for each
identified sequence value (i.e., a relative abundance of

10 each gene transcript).

        The list of identified sequence values (or genes
corresponding thereto) can then be sorted by abundance in
the cDNA population. A multitude of additional types of
comparisons or dimensions are possible.

15   For example (to be described below in greater detail),
steps (a) and (b) can be repeated for two different
libraries (sometimes referred to as a "target" library and
a "subtractant" library). Then, for each identified
sequence value (or gene transcript), a "ratio" value is

20 obtained by dividing the abundance number (for that
identified sequence value) for the target library, by the
abundance number (for that identified sequence value) for
the subtractant library.

        In fact, subtraction may be carried out on multiple

25 libraries. It is possible to add the transcripts from
several libraries (for example, three) and then to divide
them by another set of transcripts from multiple libraries
(again, for example, three). Notation for this operation
may be abbreviated as (A+B+C) / (D+E+F), where the capital

30 letters each indicate an entire library. Optionally the
abundance numbers of transcripts in the summed libraries
may be divided by the total sample size before subtraction.

        Unlike standard hybridization technology which permits
a single subtraction of two libraries, once one has

35 processed a set or library transcript sequences and stored
them in the computer, any number of subtractions can be
performed on the library. For example, by this method,
ratio values can be obtained by dividing relative abundance

values in a first library by corresponding values in a second library and vice versa.

In variations on step (a), the library consists of nucleotide sequences derived from cDNA clones. Examples of
5 databases which can be searched for areas of homology (similarity) in step (a) include the commercially available databases known as Genbank (NIH) EMBL (European Molecular Biology Labs, Germany), and GENESEQ (Intelligenetics, Mountain View, California).
10 One homology search algorithm which can be used to implement step (a) is the algorithm described in the paper by D.J. Lipman and W.R. Pearson, entitled "Rapid and Sensitive Protein Similarity Searches," Science, 227:1435 (1985). In this algorithm, the homologous regions are
15 searched in a two-step manner. In the first step, the highest homologous regions are determined by calculating a matching score using a homology score table. The parameter "Ktup" is used in this step to establish the minimum window size to be shifted for comparing two sequences. Ktup also
20 sets the number of bases that must match to extract the highest homologous region among the sequences. In this step, no insertions or deletions are applied and the homology is displayed as an initial (INIT) value.

In the second step, the homologous regions are aligned
25 to obtain the highest matching score by inserting a gap in order to add a probable deleted portion. The matching score obtained in the first step is recalculated using the homology score Table and the insertion score Table to an optimized (OPT) value in the final output.
30 DNA homologies between two sequences can be examined graphically using the Harr method of constructing dot matrix homology plots (Needleman, S.B. and Wunsch, C.O., J. Mom. Biol 48:443 (1970)). This method produces a two-dimensional plot which can be useful in determining
35 regions of homology versus regions of repetition.

However, in a class of preferred embodiments, step (a) is implemented by processing the library data in the commercially available computer program known as the INHERIT 670 Sequence Analysis System, available from

Applied Biosystems Inc. (Foster City, California),
including the software known as the Factura software (also
available from Applied Biosystems Inc.). The Factura
program preprocesses each library sequence to "edit out"
5   portions thereof which are not likely to be of interest,
such as the vector used to prepare the library. Additional
sequences which can be edited out or masked (ignored by the
search tools) include but are not limited to the polyA tail
and repetitive GAG and CCC sequences. A low-end search
10   program can be written to mask out such "low-information"
sequences, or programs such as BLAST can ignore the low-
information sequences.

In the algorithm implemented by the INHERIT 670
Sequence Analysis System, the Pattern Specification
15   Language (developed by TRW Inc.) is used to determine
regions of homology. "There are three parameters that
determine how INHERIT analysis runs sequence comparisons:
window size, window offset and error tolerance. Window
size specifies the length of the segments into which the
20   query sequence is subdivided. Window offset specifies
where to start the next segment [to be compared], counting
from the beginning of the previous segment. Error
tolerance specifies the total number of insertions,
deletions and/or substitutions that are tolerated over the
25   specified word length. Error tolerance may be set to any
integer between 0 and 6. The default settings are window
tolerance=20, window offset=10 and error tolerance=3."
INHERIT Analysis Users Manual, pp.2-15. Version 1.0,
Applied Biosystems, Inc., October 1991.
30   Using a combination of these three parameters, a
database (such as a DNA database) can be searched for
sequences containing regions of homology and the
appropriate sequences are scored with an initial value.
Subsequently, these homologous regions are examined using
35   dot matrix homology plots to determine regions of homology
versus regions of repetition. Smith-Waterman alignments
can be used to display the results of the homology search.
The INHERIT software can be executed by a Sun computer
system programmed with the UNIX operating system.

Search alternatives to INHERIT include the BLAST
program, GCG (available from the Genetics Computer Group,
WI) and the Dasher program (Temple Smith, Boston
University, Boston, MA). Nucleotide sequences can be
5   searched against Genbank, EMBL or custom databases such as
GENESEQ (available from Intelligenetics, Mountain View, CA)
or other databases for genes. In addition, we have
searched some sequences against our own in-house database.

In preferred embodiments, the transcript sequences are
10   analyzed by the INHERIT software for best conformance with
a reference gene transcript to assign a sequence identifier
and assigned the degree of homology, which together are the
identified sequence value and are input into, and further
processed by, a Macintosh personal computer (available from
15   Apple) programmed with an "abundance sort and subtraction
analysis" computer program (to be described below).

Prior to the abundance sort and subtraction analysis
program (also denoted as the "abundance sort" program),
identified sequences from the cDNA clones are assigned
20   value (according to the parameters given above) by degree
of match according to the following categories: "exact"
matches (regions with a high degree of identity),
homologous human matches (regions of high similarity, but
not "exact" matches), homologous non-human matches (regions
25   of high similarity present in species other than human), or
non matches (no significant regions of homology to
previously identified nucleotide sequences stored in the
form of the database). Alternately, the degree of match
can be a numeric value as described below.
30   With reference again to the step of identifying
matches between reference sequences and database entries,
protein and peptide sequences can be deduced from the
nucleic acid sequences. Using the deduced polypeptide
sequence, the match identification can be performed in a
35   manner analogous to that done with cDNA sequences. A
protein sequence is used as a query sequence and compared
to the previously identified sequences contained in a
database such as the Swiss/Prot, PIR and the NBRF Protein
database to find homologous proteins. These proteins are

27

initially scored for homology using a homology score Table
(Orcutt, B.C. and Dayoff, M.O. Scoring Matrices, PIR
Report MAT - 0285 (February 1985)) resulting in an INIT
score. The homologous regions are aligned to obtain the
5   highest matching scores by inserting a gap which adds a
probable deleted portion. The matching score is
recalculated using the homology score Table and the
insertion score Table resulting in an optimized (OPT)
score. Even in the absence of knowledge of the proper
10   reading frame of an isolated sequence, the above-described
protein homology search may be performed by searching all 3
reading frames.

Peptide and protein sequence homologies can also be
ascertained using the INHERIT 670 Sequence Analysis System
15   in an analogous way to that used in DNA sequence
homologies. Pattern Specification Language and parameter
windows are used to search protein databases for sequences
containing regions of homology which are scored with an
initial value. Subsequent display in a dot-matrix homology
20   plot shows regions of homology versus regions of
repetition. Additional search tools that are available to
use on pattern search databases include PLsearch Blocks
(available from Henikoff & Henikoff, University of
Washington, Seattle), Dasher and GCG. Pattern search
25   databases include, but are not limited to, Protein Blocks
(available from Henikoff & Henikoff, University of
Washington, Seattle), Brookhaven Protein (available from
the Brookhaven National Laboratory, Brookhaven, MA),
PROSITE (available from Amos Bairoch, University of Geneva,
30   Switzerland), ProDom (available from Temple Smith, Boston
University), and PROTEIN MOTIF FINGERPRINT (available from
University of Leeds, United Kingdom).

The ABI Assembler application software, part of the
INHERIT DNA analysis system (available from Applied
35   Biosystems, Inc., Foster City, CA), can be employed to
create and manage sequence assembly projects by assembling
data from selected sequence fragments into a larger
sequence. The Assembler software combines two advanced
computer technologies which maximize the ability to

assemble sequenced DNA fragments into Assemblages, a
special grouping of data where the relationships between
sequences are shown by graphic overlap, alignment and
statistical views. The process is based on the
5   Meyers-Kececioglu model of fragment assembly (INHERIT™
Assembler User's Manual, Applied Biosystems, Inc., Foster
City, CA), and uses graph theory as the foundation of a
very rigorous multiple sequence alignment engine for
assembling DNA sequence fragments. Other assembly programs
10 that can be used include MEGALIGN (available from DNASTAR
Inc., Madison, WI), Dasher and STADEN (available from Roger
Staden, Cambridge, England).

     Next, with reference to Fig. 2, we describe in more
detail the "abundance sort" program which implements above-
15 mentioned "step (b)" to tabulate the number of sequences of
· the library which match each database entry (the "abundance
number" for each database entry).

     Fig. 2 is a flow chart of a preferred embodiment of
the abundance sort program. A source code listing of this
20 embodiment of the abundance sort program is set forth in
Table 5. In the Table 5 implementation, the abundance sort
program is written using the FoxBASE programming language
commercially available from Microsoft Corporation.
Although FoxBASE was the program chosen for the first
25 iteration of this technology, it should not be considered
limiting. Many other programming languages, Sybase being a
particularly desirable alternative, can also be used, as
will be obvious to one with ordinary skill in the art. The
subroutine names specified in Fig. 2 correspond to
30 subroutines listed in Table 5.

     With reference again to Fig. 2, the "Identified
Sequences" are transcript sequences representing each
sequence of the library and a corresponding identification
of the database entry (if any) which it matches. In other
35 words, the "Identified Sequences" are transcript sequences
representing the output of above-discussed "step (a)."

     Fig. 3 is a block diagram of a system for implementing
the invention. The Fig. 3 system includes library
generation unit 2 which generates a library and asserts an

output stream of transcript sequences indicative of the
biological sequences comprising the library. Programmed
processor 4 receives the data stream output from unit 2 and
processes this data in accordance with above-discussed

5   "step (a)" to generate the Identified Sequences. Processor
4 can be a processor programmed with the commercially
available computer program known as the INHERIT 670
Sequence Analysis System and the commercially available
computer program known as the Factura program (both

10   available from Applied Biosystems Inc.) and with the UNIX
operating system.

     Still with reference to Fig. 3, the Identified
Sequences are loaded into processor 6 which is programmed
with the abundance sort program. Processor 6 generates the

15   Final Transcript sequences indicated in both Figs. 2 and 3.
Fig. 4 shows a more detailed block diagram of a planned
relational computer system, including various searching
techniques which can be implemented, along with an
assortment of databases to query against.

20   With reference to Fig. 2, the abundance sort program
first performs an operation known as "Tempnum" on the
Identified Sequences, to discard all of the Identified
Sequences except those which match database entries of
selected types. For example, the Tempnum process can

25   select Identified Sequences which represent matches of the
following types with database entries (see above for
definition): "exact" matches, human "homologous" matches,
"other species" matches representing genes present in
species other than human), "no" matches (no significant

30   regions of homology with database entries representing
previously identified nucleotide sequences), "I" matches
(Incyte for not previously known DNA sequences), or "X"
matches (matches ESTs in reference database). This
eliminates the U, S, M, V, A, R and D sequence (see Table 1

35   for definitions).

     The identified sequence values selected during the
"Tempnum" process then undergo a further selection (weeding
out) operation known as "Tempred." This operation can, for

exampl , discard all identified sequence values
representing matches with selected database entries.

The identified sequence values selected during the
"Tempred" process are then classified according to library,
5   during the "Tempdesig" operation. It is contemplated that
the "Identified Sequences" can represent sequences from a
single library, or from two or more libraries.

Consider first the case that the identified sequence
values represent sequences from a single library. In this
10 case, all the identified sequence values determined during
"Tempred" undergo sorting in the "Templib" operation,
further sorting in the "Libsort" operation, and finally
additional sorting in the "Temptarsort" operation. For
example, these three sorting operations can sort the
15 identified sequences in order of decreasing "abundance
number" (to generate a list of decreasing abundance
numbers, each abundance number corresponding to a unique
identified sequence entry, or several lists of decreasing
abundance numbers, with the abundance numbers in each list
20 corresponding to database entries of a selected type) with
redundancies eliminated from each sorted list. In this
case, the operation identified as "Cruncher" can be
bypassed, so that the "Final Data" values are the organized
transcript sequences produced during the "Temptarsort"
25 operation.

We next consider the case that the transcript
sequences produced during the "Tempred" operation represent
sequences from two libraries (which we will denote the
"target" library and the "subtractant" library). For
30 example, the target library may consist of cDNA sequences
from clones of a diseased cell, while the subtractant
library may consist of cDNA sequences from clones of the
diseased cell after treatment by exposure to a drug. For
another example, the target library may consist of cDNA
35 sequences from clones of a cell type from a young human,
while the subtractant library may consist of cDNA sequences
from clones of the same cell type from the same human at
different ages.

In this case, the "Tempdesig" operation routes all transcript sequences representing the target library for processing in accordance with "Templib" (and then "Libsort" and "Temptarsort"), and routes all transcript sequences

5   representing the subtractant library for processing in accordance with "Tempsub" (and then "Subsort" and "Tempsubsort"). For example, the consecutive "Templib," "Libsort," and "Temptarsort" sorting operations sort identified sequences from the target library in order of

10 decreasing abundance number (to generate a list of decreasing abundance numbers, each abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected

15 type) with redundancies eliminated from each sorted list. The consecutive "Tempsub," "Subsort," and "Tempsubsort" sorting operations sort identified sequences from the subtractant library in order of decreasing abundance number (to generate a list of decreasing abundance numbers, each

20 abundance number corresponding to a database entry, or several lists of decreasing abundance numbers, with the abundance numbers in each list corresponding to database entries of a selected type) with redundancies eliminated from each sorted list.

25   The transcript sequences output from the "Temptarsort" operation typically represent sorted lists from which a histogram could be generated in which position along one (e.g., horizontal) axis indicates abundance number (of target library sequences), and position along another

30 (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type). Similarly, the transcript sequences output from the "Tempsubsort" operation typically represent sorted lists from which a histogram could be generated in which position along one

35 (e.g., horizontal) axis indicates abundance number (of subtractant library sequences), and position along another (e.g., vertical) axis indicates identified sequence value (e.g., human or non-human gene type).

The transcript sequences (sorted lists) output from
the Tempsubsort and Temptarsort sorting operations are
combined during the operation identified as "Cruncher."
The "Cruncher" process identifies pairs of corresponding
5   target and subtractant abundance numbers (both representing
the same identified sequence value), and divides one by the
other to generate a "ratio" value for each pair of
corresponding abundance numbers, and then sorts the ratio
values in order of decreasing ratio value. The data output
10   from the "Cruncher" operation (the Final Transcript
sequence in Fig. 2) is typically a sorted list from which a
histogram could be generated in which position along one
axis indicates the size of a ratio of abundance numbers
(for corresponding identified sequence values from target
15   and subtractant libraries) and position along another axis
indicates identified sequence value (e.g., gene type).
        Preferably, prior to obtaining a ratio between the two
library abundance values, the Cruncher operation also
divides each ratio value by the total number of sequences
20   in one or both of the target and subtractant libraries.
The resulting lists of "relative" ratio values generated by
the Cruncher operation are useful for many medical,
scientific, and industrial applications. Also preferably,
the output of the Cruncher operation is a set of lists,
25   each list representing a sequence of decreasing ratio
values for a different selected subset (e.g. protein
family) of database entries.
        In one example, the abundance sort program of the
invention tabulates for a library the numbers of mRNA
30   transcripts corresponding to each gene identified in a
database. These numbers are divided by the total number of
clones sampled. The results of the division reflect the
relative abundance of the mRNA transcripts in the cell type
or tissue from which they were obtained. Obtaining this
35   final data set is referred to herein as "gene transcript
image analysis." The resulting subtracted data show
exactly what proteins and genes are upregulated and
downregulated in highly detailed complexity.

6.6.   HUVEC cDNA LIBRARY

Table 2 is an abundance table listing the various gene transcripts in an induced HUVEC library. The transcripts are listed in order of decreasing abundance. This
5   computerized sorting simplifies analysis of the tissue and speeds identification of significant new proteins which are specific to this cell type. This type of endothelial cell lines tissues of the cardiovascular system, and the more that is known about its composition, particularly in
10   response to activation, the more choices of protein targets become available to affect in treating disorders of this tissue, such as the highly prevalent atherosclerosis.

6.7.   MONOCYTE-CELL AND MAST-CELL cDNA LIBRARIES

Tables 3 and 4 show truncated comparisons of two
15   libraries. In Tables 3 and 4 the "normal monocytes" are the HMC-1 cells, and the "activated macrophages" are the THP-1 cells pretreated with PMA and activated with LPS. Table 3 lists in descending order of abundance the most abundant gene transcripts for both cell types. With only
20   15 gene transcripts from each cell type, this table permits quick, qualitative comparison of the most common transcripts. This abundance sort, with its convenient side-by-side display, provides an immediately useful research tool. In this example, this research tool
25   discloses that 1) only one of the top 15 activated macrophage transcripts is found in the top 15 normal monocyte gene transcripts (poly A binding protein); and 2) a new gene transcript (previously unreported in other databases) is relatively highly represented in activated
30   macrophages but is not similarly prominent in normal macrophages. Such a research tool provides researchers with a short-cut to new proteins, such as receptors, cell-surface and intracellular signalling molecules, which can serve as drug targets in commercial drug screening
35   programs. Such a tool could save considerable time over that consumed by a hit and miss discovery program aimed at identifying important proteins in and around cells, because those proteins carrying out everyday cellular functions and

represented as steady state mRNA are quickly eliminated from further characterization.

This illustrates how the gene transcript profiles change with altered cellular function. Those skilled in the art know that the biochemical composition of cells also changes with other functional changes such as cancer, including cancer's various stages, and exposure to toxicity. A gene transcript subtraction profile such as in Table 3 is useful as a first screening tool for such gene expression and protein studies.

6.8. SUBTRACTION ANALYSIS OF NORMAL MONOCYTE-CELL AND ACTIVATED MONOCYTE CELL cDNA LIBRARIES

Once the cDNA data are in the computer, the computer program as disclosed in Table 5 was used to obtain ratios of all the gene transcripts in the two libraries discussed in Example 6.7, and the gene transcripts were sorted by the descending values of their ratios. If a gene transcript is not represented in one library, that gene transcript's abundance is unknown but appears to be less than 1. As an approximation -- and to obtain a ratio, which would not be possible if the unrepresented gene were given an abundance of zero -- genes which are represented in only one of the two libraries are assigned an abundance of 1/2. Using 1/2 for unrepresented clones increases the relative importance of "turned-on" and "turned-off" genes, whose products would be drug candidates. The resulting print-out is called a subtraction table and is an extremely valuable screening method, as is shown by the following data.

Table 4 is a subtraction table, in which the normal monocyte library was electronically "subtracted" from the activated macrophage library. This table highlights most effectively the changes in abundance of the gene transcripts by activation of macrophages. Even among the first 20 gene transcripts listed, there are several unknown gene transcripts. Thus, electronic subtraction is a useful tool with which to assist researchers in identifying much more quickly the basic biochemical changes between two cell types. Such a tool can save universities and pharmaceutical companies which spend billions of dollars on

research valuable time and laboratory resources at the
early discovery stage and can speed up the drug development
cycle, which in turn permits researchers to set up drug
screening programs much earlier. Thus, this research tool

5 provides a way to get new drugs to the public faster and
more economically.

Also, such a subtraction table can be obtained for
patient diagnosis. An individual patient sample (such as
monocytes obtained from a biopsy or blood sample) can be

10 compared with data provided herein to diagnose conditions
associated with macrophage activation.

Table 4 uncovered many new gene transcripts (labeled
Incyte clones). Note that many genes are turned on in the
activated macrophage (i.e., the monocyte had a 0 in the

15 bgfreq column). This screening method is superior to other
screening techniques, such as the western blot, which are
incapable of uncovering such a multitude of discrete new
gene transcripts.

The subtraction-screening technique has also uncovered

20 a high number of cancer gene transcripts (oncogenes rho,
ETS2, rab-2 ras, YPT1-related, and acute myeloid leukemia
mRNA) in the activated macrophage. These transcripts may
be attributed to the use of immortalized cell lines and are
inherently interesting for that reason. This screening

25 technique offers a detailed picture of upregulated
transcripts including oncogenes, which helps explain why
anti-cancer drugs interfere with the patient's immunity
mediated by activated macrophages. Armed with knowledge
gained from this screening method, those skilled in the art

30 can set up more targeted, more effective drug screening
programs to identify drugs which are differentially
effective against 1) both relevant cancers and activated
macrophage conditions with the same gene transcript
profile; 2) cancer alone; and 3) activated macrophage

35 conditions.

Smooth muscle senescent protein (22 kd) was
upregulated in the activated macrophage, which indicates
that it is a candidate to block in controlling
inflammation.

6.9. SUBTRACTION ANALYSIS OF NORMAL LIVER CELLS AND HEPATITIS INFECTED LIVER CELL cDNA LIBRARIES

In this example, rats are exposed to hepatitis virus and maintained in the colony until they show definite signs

5 of hepatitis. Of the rats diagnosed with hepatitis, one half of the rats are treated with a new anti-hepatitis agent (AHA). Liver samples are obtained from all rats before exposure to the hepatitis virus and at the end of AHA treatment or no treatment. In addition, liver samples

10 can be obtained from rats with hepatitis just prior to AHA treatment.

The liver tissue is treated as described in Examples 6.2 and 6.3 to obtain mRNA and subsequently to sequence cDNA. The cDNA from each sample are processed and analyzed

15 for abundance according to the computer program in Table 5. The resulting gene transcript images of the cDNA provide detailed pictures of the baseline (control) for each animal and of the infected and/or treated state of the animals. cDNA data for a group of samples can be combined into a

20 group summary gene transcript profile for all control samples, all samples from infected rats and all samples from AHA-treated rats.

Subtractions are performed between appropriate individual libraries and the grouped libraries. For

25 individual animals, control and post-study samples can be subtracted. Also, if samples are obtained before and after AHA treatment, that data from individual animals and treatment groups can be subtracted. In addition, the data for all control samples can be pooled and averaged. The

30 control average can be subtracted from averages of both post-study AHA and post-study non-AHA cDNA samples. If pre- and post-treatment samples are available, pre- and post-treatment samples can be compared individually (or electronically averaged) and subtracted.

35 These subtraction tables are used in two general ways. First, the differences are analyzed for gene transcripts which are associated with continuing hepatic deterioration or healing. The subtraction tables are tools to isolate the effects of the drug treatment from the underlying basic

40 pathology of hepatitis. Because hepatitis affects many

parameters, additional liver toxicity has been difficult to detect with only blood tests for the usual enzymes. The gene transcript profile and subtraction provides a much more complex biochemical picture which researchers have
5 needed to analyze such difficult problems.

Second, the subtraction tables provide a tool for identifying clinical markers, individual proteins or other biochemical determinants which are used to predict and/or evaluate a clinical endpoint, such as disease, improvement
10 due to the drug, and even additional pathology due to the drug. The subtraction tables specifically highlight genes which are turned on or off. Thus, the subtraction tables provide a first screen for a set of gene transcript candidates for use as clinical markers. Subsequently,
15 electronic subtractions of additional cell and tissue libraries reveal which of the potential markers are in fact found in different cell and tissue libraries. Candidate gene transcripts found in additional libraries are removed from the set of potential clinical markers. Then, tests of
20 blood or other relevant samples which are known to lack and have the relevant condition are compared to validate the selection of the clinical marker. In this method, the particular physiologic function of the protein transcript need not be determined to qualify the gene transcript as a
25 clinical marker.

6.10. ELECTRONIC NORTHERN BLOT

One limitation of electronic subtraction is that it is difficult to compare more than a pair of images at once. Once particular individual gene products are identified as
30 relevant to further study (via electronic subtraction or other methods), it is useful to study the expression of single genes in a multitude of different tissues. In the lab, the technique of "Northern" blot hybridization is used for this purpose. In this technique, a single cDNA, or a
35 probe corresponding thereto, is labeled and then hybridized against a blot containing RNA samples prepared from a multitude of tissues or cell types. Upon autoradiography,

the pattern of expression of that particular gene, one at a
time, can be quantitated in all the included samples.

In contrast, a further embodiment of this invention is
the computerized form of this process, termed here

5   "electronic northern blot." In this variation, a single
gene is queried for expression against a multitude of
prepared and sequenced libraries present within the
database. In this way, the pattern of expression of any
single candidate gene can be examined instantaneously and

10   effortlessly. More candidate genes can thus be scanned,
leading to more frequent and fruitfully relevant
discoveries. The computer program included as Table 5
includes a program for performing this function, and Table
6 is a partial listing of entries of the database used in

15   the electronic northern blot analysis.

6.11.   PHASE I CLINICAL TRIALS

Based on the establishment of safety and effectiveness
in the above animal tests, Phase I clinical tests are
undertaken. Normal patients are subjected to the usual

20   preliminary clinical laboratory tests. In addition,
appropriate specimens are taken and subjected to gene
transcript analysis. Additional patient specimens are
taken at predetermined intervals during the test. The
specimens are subjected to gene transcript analysis as

25   described above. In addition, the gene transcript changes
noted in the earlier rat toxicity study are carefully
evaluated as clinical markers in the followed patients.
Changes in the gene transcript analyses are evaluated as
indicators of toxicity by correlation with clinical signs

30   and symptoms and other laboratory results. In addition,
subtraction is performed on individual patient specimens
and on averaged patient specimens. The subtraction
analysis highlights any toxicological changes in the
treated patients. This is a highly refined determinant of

35   toxicity. The subtraction method also annotates clinical
markers. Further subgroups can be analyzed by subtraction
analysis, including, for example, 1) segregation by

occurrence and type of adverse effect; and 2) segregation
by dosage.

6.12. GENE TRANSCRIPT IMAGING ANALYSIS IN CLINICAL STUDIES

A gene transcript imaging analysis (or multiple gene
5   transcript imaging analyses) is a useful tool in other
clinical studies. For example, the differences in gene
transcript imaging analyses before and after treatment can
be assessed for patients on placebo and drug treatment.
This method also effectively screens for clinical markers
10   to follow in clinical use of the drug.

6.13. COMPARATIVE GENE TRANSCRIPT ANALYSIS BETWEEN SPECIES

The subtraction method can be used to screen cDNA
libraries from diverse sources. For example, the same cell
types from different species can be compared by gene
15   transcript analysis to screen for specific differences,
such as in detoxification enzyme systems. Such testing
aids in the selection and validation of an animal model for
the commercial purpose of drug screening or toxicological
testing of drugs intended for human or animal use. When
20   the comparison between animals of different species is
shown in columns for each species, we refer to this as an
interspecies comparison, or zoo blot.

Embodiments of this invention may employ databases
such as those written using the FoxBASE programming
25   language commercially available from Microsoft Corporation.
Other embodiments of the invention employ other databases,
such as a random peptide database, a polymer database, a
synthetic oligomer database, or a oligonucleotide database
of the type described in U.S. Patent 5,270,170, issued
30   December 14, 1993 to Cull, et al., PCT International
Application Publication No. WO 9322684, published November
11, 1993, PCT International Application Publication No. WO
9306121, published April 1, 1993, or PCT International
Application Publication No. WO 9119818, published December
35   26, 1991. These four references (whose text is
incorporated herein by reference) include teaching which

may be applied in implementing such other embodiments of
the present invention.

All references referred to in the preceding text are
hereby expressly incorporated by reference herein.

5   Various modifications and variations of the described
method and system of the invention will be apparent to
those skilled in the art without departing from the scope
and spirit of the invention. Although the invention has
been described in connection with specific preferred

10   embodiments, it should be understood that the invention as
claimed should not be unduly limited to such specific
embodiments.

TABLE 1

Designations (D)

E = Exact
H = Homologous
O = Other species
N = No match
D = Noncoding gene
U = Nonreadable
R = Repetitive DNA
A = Poly-A only
V = Vector only
M = Mitochondrial DNA
S = Skip
I = Match Incyte clone
X = EST match

Distribution (F)

C = Non-specific
P = Cell/tissue specific
U = Unknown

Species (S)

H = Human
A = Ape
P = Pig
D = Dog
V = Bovine
B = Rabbit
R = Rat
M = Mouse
S = Hamster
C = Chicken
F = Amphibian
I = Invertebrate
Z = Protozoan
G = Fungi

Library (L)

U = U937
M = HMC
T = THP-1
H = HUVEC
S = Spleen
L = Lung
Y = T & B cell
A = Adenoid

Localization (Z)

N = Nuclear
C = Cytoplasmic
K = Cytoskeleton
E = Cell surface
Z = Intracellular memb
M = Mitochondrial
S = Secreted
U = Unknown
X = Other

Status (I)

0 = No current interest
1 = Do primary analysis
2 = Primary analysis done
3 = Full length sequence
4 = Secondary analysis
5 = Tissue northern
6 = Obtain full length

Function (R)

T = Translation
L = Protein processing
R = Ribosomal protein
O = Oncogene
  G = GTP binding ptn
  V = Viral element
  Y = Kinase/phosphatase
  A = Tumor antigen related
I = Binding proteins
  D = NA-binding /transcription
  B = Surface molecule/receptor
  C = Ca^{++} binding protein
  S = Ligands/effectors
H = Stress response protein
E = Enzyme
  F = Ferroprotein
  P = Protease/inhibitor
  Z = Oxidative phosphorylation
  Q = Sugar metabolism
  M = Amino acid metabolism
  N = Nucleic acid metabolism
  W = Lipid metabolism
  K = Structural
  X = Other
  U = unknown

4 2

TABLE 2

Clone numbers 15000 through 20000
Libraries: HUVEC
Arranged by ABUNDANCE
Total clones analyzed: 5000

319 genes, for a total of 1713 Clones

	number	N	c	entry	s	descriptor
1	15365	67		HSRPL41		Riboptn L41
2	15004	65		NCYO15004		INCYTE 015004
3	15638	63		NCYO15638		INCYTE 015638
4	15390	50		NCYO15390		INCYTE 015390
5	15193	47		HSFIB1		Fibronectin
6	15220	47		RRRPL9	R	Riboptn L9
7	15280	47		NCYO15280		INCYTE 015280
8	15583	33		M62060		EST HHCH09 (IGR)
9	15662	31		HSACTCGR		Actin, gamma
10	15026	29		NCYO15026		INCYTE 015026
11	15279	24		HSEF1AR		Elf 1-alpha
12	15027	23		NCYO15027		INCYTE 015027
13	15033	20		NCYO15033		INCYTE 015033
14	15198	20		NCYO15198		INCYTE 015198
15	15809	20		HSCOLL1		Collagenase
16	15221	19		NCYO15221		INCYTE 015221
17	15263	19		NCYO15263		INCYTE 015263
18	15290	19		NCYO15290		INCYTE 015290
19	15350	18		NCYO15350		INCYTE 015350
20	15030	17		NCYO15030		INCYTE 015030
21	15234	17		NCYO15234		INCYTE 015234
22	15459	16		NCYO15459		INCYTE 015459
23	15353	15		NCYO15353		INCYTE 015353
24	15378	15		S76965		Ptn kinase inhib
25	15255	14		HUMTHYB4		Thymosin beta-4
26	15401	14		HSLIPCR		Lipocortin I
27	15425	14		HSPOLYAB		Poly-A bp
28	18212	14		HUMTHYMA		Thymosin, alpha
29	18216	14		HSMRP1		Motility relat ptn; MRP-1;CD-9
30	15189	13		HS18D		Interferon induc ptn 1-8D
31	15031	12		HUMFKBP		FK506 bp
32	15306	12		HSH2AZ		Histone H2A
33	15621	12		HUMLEC		Lectin, B-galbp, 14kDa
34	15789	11		NCYO15789		INCYTE 015789
35	16578	11		HSRPS11		Riboptn S11
36	16632	11		M61984		EST HHCA13 (IGR)
37	18314	11		NCYO18314		INCYTE 018314
38	15367	10		NCYO15367		INCYTE 015367
39	15415	10		HSIFNIN1		interferon induc mRNA
40	15633	10		HSLDHAR		Lactate dehydrogenase
41	15813	10		CHKNMHCB		C Myosin heavy chain B
42	18210	10		NCYO18210		INCYTE 018210
43	18233	10		HSRPII140		RNA polymerase II
44	18996	10		NCYO18996		INCYTE 018996
45	15088	9		HUMFERL		Ferritin, light chain
46	15714	9		NCYO15714		INCYTE 015714
47	15720	9		NCYO15720		INCYTE 015720
48	15863	9		NCYO15863		INCYTE 015863
49	16121	9		HSET		Endothelin
50	18252	9		NCYO18252		INCYTE 018252
51	15351	8		HUMALBP		Lipid bp, adipocyte
52	15370	8		NCYO15370		INCYTE 015370

TABLE 2 Con't

	number	N	c	entry	s	descriptor
53	15670	8		BTCIASHI	V	NADH-ubiq oxidoreductase
54	15795	8		NCY015795		INCYTE 015795
55	16245	8		NCY016245		INCYTE 016245
56	18262	8		NCY018262		INCYTE 018262
57	18321	8		HSRPL17		Riboptn L17
58	15126	7		XLRPL1BRF		Riboptn L1
59	15133	7		HSAC07		Actin, beta
60	15245	7		NCY015245		INCYTE 015245
61	15288	7		NCY015288		INCYTE 015288
62	15294	7		HSGAPDR		G-3-PD
63	15442	7		HUMLAMB		Laminin receptor, 54kDa
64	15485	7		HSNGMRNA		Uracil DNA glycosylase
65	16646	7		NCY016646		INCYTE 016646
66	18003	7		HUMPAIA		Plsmnogen activ gene
67	15032	6		HUMUB		Ubiquitin
68	15267	6		HSRPS8		Riboptn S8
69	15295	6		NCY015295		INCYTE 015295
70	15458	6		RNRPS10R	R	Riboptn S10
71	15832	6		RSGALEM	R	UDP-galactose epimerase
72	15928	6		HUMAPOJ		Apolipoptn J
73	16598	6		HUMTBBM40		Tubulin, beta
74	18218	6		NCY018218		INCYTE 018218
75	18499	6		HSP27		Hydrophobic ptn p27
76	18963	6		NCY018963		INCYTE 018963
77	18997	6		NCY018997		INCYTE 018997
78	15432	5		HSAGALAR		Galactosidase A, alpha
79	15475	5		NCY015475		INCYTE 015475
80	15721	5		NCY015721		INCYTE 015721
81	15865	5		NCY015865		INCYTE 015865
82	16270	5		NCY016270		INCYTE 016270
83	16886	5		NCY016886		INCYTE 016886
84	18500	5		NCY018500		INCYTE 018500
85	18503	5		NCY018503		INCYTE 018503
86	19672	5		RRRPL34	R	Riboptn L34
87	15086	4		XLRPL1AR	F	Riboptn L1a
88	15113	4		HUMIFNWRS		tRNA synthetase, trp
89	15242	4		NCY015242		INCYTE 015242
90	15249	4		NCY015249		INCYTE 015249
91	15377	4		NCY015377		INCYTE 015377
92	15407	4		NCY015407		INCYTE 015407
93	15473	4		NCY015473		INCYTE 015473
94	15588	4		HSRPS12		Riboptn S12
95	15684	4		HSEF1G		Elf 1-gamma
96	15782	4		NCY015782		INCYTE 015782
97	15916	4		HSRPS18		Riboptn S18
98	15930	4		NCY015930		INCYTE 015930
99	16108	4		NCY016108		INCYTE 016108
100	16133	4		NCY016133		INCYTE 016133

NORMAL MONONCYTE VS. ACTIVATED MACROPHAGE

Top 15 Most Abundant Genes

NORMAL

1   Elongation factor-1 alpha
2   Ribosomal phosphoprotein
3   Ribosomal protein S8 homolog
4   Beta-Globin
5   Ferritin H chain
6   Ribosomal protein L7
7   Nucleoplasmin
8   Ribosomal protein S20 homolog
9   Transferrin receptor
10 Poly-A binding protein
11 Translationally controlled tumor ptn
12 Ribosomal protein S25
13 Signal recognition particle SRP9
14 Histone H2A.Z
15 Ribosomal protein Ke-3

ACTIVATED

Interleukin-1 beta
Macrophage inflammatory protein-1
Interleukin-8
Lymphocyte activation gene
Elongation factor-1 alpha
Beta actin
Rantes T-cell specific protein
Poly A binding protein
Osteopontin; nephropontin
Tumor Necrosis Factor-alpha
INCYTE clone 011050
Cu/Zn superoxide dismutase
Adenylate cyclase (yeast homolog)
NGF-related B cell activation molecule
Protease Nexin-1, glial-derived

TABLE 3

4 5

TABLE 4

Libraries: THP-1
Subtracting: HMC
Sorted by ABUNDANCE
Total clones analyzed: 7375

1057 genes, for a total of 2151 clones

number	entry	s	descriptor	bgfreq	rfend	ratio
10022	HUMIL1		IL 1-beta	0	131	262.00
10036	HSMDNCF		IL-8	0	119	238.00
10089	HSLAG1CDN		Lymphocyte activ gene	0	71	142.00
10060	HUMTCSM		RANTES	0	23	46.000
10003	HUMMIP1A		MIP-1	3	121	40.333
10689	HSOP		Osteopontin	0	20	40.000
11050	NCY011050		INCYTE 011050	0	17	34.000
10937	HSTNFR		TNF-alpha	0	17	34.000
10176	HSSOD		Superoxide dismutase	0	14	28.000
10886	HSCDW40		B-cell activ,NGF-relat	0	10	20.000
10186	HUMAPR		Early resp PMA-induc	0	9	18.000
10967	HUMGDN		PN-1, glial-deriv	0	9	18.000
11353	NCY011353		INCYTE 011353	0	8	16.000
10298	NCY010298		INCYTE 010298	0	7	14.000
10215	HUM4COLA		Collagenase, type IV	0	6	12.000
10276	NCY010276		INCYTE 010276	0	6	12.000
10488	NCY010488		INCYTE 010488	0	6	12.000
11138	NCY011138		INCYTE 011138	0	6	12.000
10037	HUMCAPPRO		Adenylate cyclase	1	10	10.000
10840	HUMADCY		Adenylate cyclase	0	5	10.000
10672	HSCD44E		Cell adhesion glptn	0	5	10.000
12837	HUMCYCLOX		Cyclooxygenase-2	0	5	10.000
10001	NCY010001		INCYTE 010001	0	5	10.000
10005	NCY010005		INCYTE 010005	0	5	10.000
10294	NCY010294		INCYTE 010294	0	5	10.000
10297	NCY010297		INCYTE 010297	0	5	10.000
10403	NCY010403		INCYTE 010403	0	5	10.000
10699	NCY010699		INCYTE 010699	0	5	10.000
10966	NCY010966		INCYTE 010966	0	5	10.000
12092	NCY012092		INCYTE 012092	0	5	10.000
12549	HSRHOB		Oncogene rho	0	5	10.000
10691	HUMARF1BA		ADP-ribosylation fctr	0	4	8.000
12106	HSADSS		Adenylosuccinate synthetase	0	4	8.000
10194	HSCATHL		Cathepsin L	0	4	8.000
10479	CLMCYCA	I	Cyclin A	0	4	8.000
10031	NCY010031		INCYTE 010031	0	4	8.000
10203	NCY010203		INCYTE 010203	0	4	8.000
10288	NCY010288		INCYTE 010288	0	4	8.000
10372	NCY010372		INCYTE 010372	0	4	8.000
10471	NCY010471		INCYTE 010471	0	4	8.000
10484	NCY010484		INCYTE 010484	0	4	8.000
10859	NCY010859		INCYTE 010859	0	4	8.000
10890	NCY010890		INCYTE 010890	0	4	8.000
11511	NCY011511		INCYTE 011511	0	4	8.000
11868	NCY011868		INCYTE 011868	0	4	8.000
12820	NCY012820		INCYTE 012820	0	4	8.000
10133	HSI1RAP		IL-1 antagonist	0	4	8.000
10516	HUMP2A		Phosphatase, regul 2A	0	4	8.000
11063	HUMB94		TNF-induc response	0	4	8.000
11140	HSHB15RNA		HB15 gene; new Ig	0	3	6.000
10788	NCY001713		INCYTE 001713	0	3	6.000
10033	NCY010033		INCYTE 010033	0	3	6.000
10035	NCY010035		INCYTE 010035	0	3	6.000
10084	NCY010084		INCYTE 010084	0	3	6.000
10236	NCY010236		INCYTE 010236	0	3	6.000
10383	NCY010383		INCYTE 010383	0	3	6.000

4 6

TABLE 4 Con't

number	entry	s descriptor	bgfreq	rfend	ratio
10450	NCY010450	INCYTE 010450	0	3	6.000
10470	NCY010470	INCYTE 010470	0	3	6.000
10504	NCY010504	INCYTE 010504	0	3	6.000
10507	NCY010507	INCYTE 010507	0	3	6.000
10598	NCY010598	INCYTE 010598	0	3	6.000
10779	NCY010779	INCYTE 010779	0	3	6.000
10909	NCY010909	INCYTE 010909	0	3	6.000
10976	NCY010976	INCYTE 010976	0	3	6.000
10985	NCY010985	INCYTE 010985	0	3	6.000
11052	NCY011052	INCYTE 011052	0	3	6.000
11068	NCY011068	INCYTE 011068	0	3	6.000
11134	NCY011134	INCYTE 011134	0	3	6.000
11136	NCY011136	INCYTE 011136	0	3	6.000
11191	NCY011191	INCYTE 011191	0	3	6.000
11219	NCY011219	INCYTE 011219	0	3	6.000
11386	NCY011386	INCYTE 011386	0	3	6.000
11403	NCY011403	INCYTE 011403	0	3	6.000
11460	NCY011460	INCYTE 011460	0	3	6.000
11618	NCY011618	INCYTE 011618	0	3	6.000
11686	NCY011686	INCYTE 011686	0	3	6.000
12021	NCY012021	INCYTE 012021	0	3	6.000
12025	NCY012025	INCYTE 012025	0	3	6.000
12320	NCY012320	INCYTE 012320	0	3	6.000
12330	NCY012330	INCYTE 012330	0	3	6.000
12853	NCY012853	INCYTE 012853	0	3	6.000
14386	NCY014386	INCYTE 014386	0	3	6.000
14391	NCY014391	INCYTE 014391	0	3	6.000

TABLE 5

```
* Master menu for SUBTRACTION output
SET TALK OFF
SET SAFETY OFF
SET EXACT ON
SET TYPEAHEAD TO 0
CLEAR
SET DEVICE TO SCREEN
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE '           ' TO Target1
STORE '           ' TO Target2
STORE '           ' TO Target3
STORE.'           ' TO Object1
STORE '           ' TO Object2
STORE '           ' TO Object3
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO PTF
STORE 1 TO BAIL
DO WHILE .T.
* Program.: Subtraction 2.fmt
* Date....: 10/11/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes...: Format file Subtraction 2
*
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',9 COLOR 0,0,0,
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,24610,-1,8947
@ PIXELS 27,134 SAY 'Subtraction Menu' STYLE 65536 FONT 'Geneva',274 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT 'Chicago',12 PICTURE '@*C Exact ' SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT 'Chicago',12 PICTURE '@*C Homologous' SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT 'Chicago',12 PICTURE '@*C Other spc' SIZE 15,84
@ PIXELS 90,152 SAY 'Matches:' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT 'Chicago',12 PICTURE '@*C Incyte' SIZE 15,65 CO
@ PIXELS 252,137 GET initiate STYLE 0 FONT 'Geneva',12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,236 GET terminate STYLE 0 FONT 'Geneva',12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,35 SAY 'Include clones ' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 252,215 SAY '->' STYLE 65536 FONT 'Geneva',14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT 'Chicago',12 PICTURE '@*C Print to file' SIZE 15,9
@ PIXELS 90,9 TO 181,109 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 90,288 TO 181,397 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 81,296 SAY 'Background:' STYLE 65536 FONT 'Geneva',270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,135 GET ANAL STYLE 65536 FONT 'Chicago',12 PICTURE '@*R Overall;Function' SIZE 4
@ PIXELS 81,26 SAY 'Target:' STYLE 65536 FONT 'Geneva',270 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,20 GET target1 STYLE 0 FONT 'Geneva',9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,20 GET target2 STYLE 0 FONT 'Geneva',9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,20 GET target3 STYLE 0 FONT 'Geneva',9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 108,299 GET object1 STYLE 0 FONT 'Geneva',9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 135,299 GET object2 STYLE 0 FONT 'Geneva',9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 162,299 GET object3 STYLE 0 FONT 'Geneva',9 SIZE 12,79 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 276,324 GET Bail STYLE 65536 FONT 'Chicago',12 PICTURE '@*R Run;Bail out' SIZE 4112
*
* EOF: Subtraction.2.fmt
READ
   IF Bail=2
   CLEAR
   CLOSE DATABASES
   USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
   SET SAFETY ON
   SCREEN 1 OFF
   RETURN
```

```
   ENDIF
STORE VAL(SYS(2)) TO STARTIME
STORE UPPER(Target1) TO Target1
STORE UPPER(Target2) TO Target2
STORE UPPER(Target3) TO Target3
STORE UPPER(Object1) TO Object1
STORE UPPER(Object2) TO Object2
STORE UPPER(Object3) TO Object3
clear
SET TALK ON
GAP = TERMINATE-INITIATE+1
GO INITIATE
COPY NEXT GAP FIELDS NUMBER,library,D,P,Z,R,ENTRY,S,DESCRIPTOR,START,RFEND,I TO TEMPNUM
USE TEMPNUM
COUNT TO TOT
COPY TO TEMPRED FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='I'
USE TEMPRED

IF Ematch=0 .AND. Pmatch=0 .AND. Omatch=0 .AND. IMATCH=0
COPY TO TEMPDESIG
ELSE
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
   IF Ematch=1
   APPEND FROM TEMPNUM FOR D='E'
   ENDIF
   IF Hmatch=1
   APPEND FROM TEMPNUM FOR D='H'
   ENDIF
   IF Omatch=1
   APPEND FROM TEMPNUM FOR D='O'
   ENDIF
   IF Imatch=1
   APPEND FROM TEMPNUM FOR D='I'.OR.D='X'
*.OR.D='N'
   ENDIF
ENDIF
COUNT TO STARTOT

COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
   APPEND FROM TEMPDESIG FOR library=UPPER(target1)
   IF target2<>'          '
   APPEND FROM TEMPDESIG FOR library=UPPER(target2)
   ENDIF
   IF target3<>'          '
   APPEND FROM TEMPDESIG FOR library=UPPER(target3)
   ENDIF
COUNT TO ANALTOT

USE TEMPDESIG
COPY STRUCTURE TO TEMPSUB
USE TEMPSUB
   APPEND FROM TEMPDESIG FOR library=UPPER(Object1)
   IF target2<>'          '
   APPEND FROM TEMPDESIG FOR library=UPPER(Object2)
   ENDIF
   IF target3<>'          '
   APPEND FROM TEMPDESIG FOR library=UPPER(Object3)
   ENDIF
COUNT TO SUBTRACTOT
SET TALK OFF
*************************************************************************************
* COMPRESSION SUBROUTINE A
? 'COMPRESSING QUERY LIBRARY'
USE TEMPLIB
```

```
SORT ON ENTRY,NUMBER TO LIBSORT
USE LIBSORT
COUNT TO IDGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= IDGENE
   PACK
   COUNT TO AUNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
   IF TESTA = TESTB.AND.DESIGA=DESIGB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO.TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPTARSORT
USE TEMPTARSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPTARCO
*****************************************************************************
* COMPRESSION SUBROUTINE B
? 'COMPRESSING TARGET LIBRARY'
USE TEMPSUB
SORT ON ENTRY,NUMBER TO SUBSORT
USE SUBSORT
COUNT TO SUBGENE
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= SUBGENE
   PACK
   COUNT TO BUNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
STORE D TO DESIGA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
STORE D TO DESIGB
   IF TESTA = TESTB.AND.DESIGA=DESIGB
```

```
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
SORT ON RFEND/D,NUMBER TO TEMPSUBSORT
USE TEMPSUBSORT
*REPLACE ALL START WITH RFEND/IDGENE*10000
COUNT TO TEMPSUBCO
*********************************************************************************
*FUSION ROUTINE
? 'SUBTRACTING LIBRARIES'
USE SUBTRACTION
COPY STRUCTURE TO CRUNCHER
SELECT 2
USE TEMPSUBSORT
SELECT 1
USE CRUNCHER
APPEND FROM TEMPTARSORT
COUNT TO BAILOUT
MARK = 0

DO WHILE .T.
SELECT 1
MARK = MARK+1
   IF MARK>BAILOUT
   EXIT
   ENDIF
GO MARK
STORE ENTRY TO SCANNER
SELECT 2
LOCATE FOR ENTRY=SCANNER
IF FOUND()
STORE RFEND TO BIT1
STORE RFEND TO BIT2
ELSE
STORE 1/2 TO BIT1
STORE 0 TO BIT2
ENDIF
SELECT 1
REPLACE BGFREQ WITH BIT2
REPLACE ACTUAL WITH BIT1
LOOP
ENDDO

SELECT 1
REPLACE ALL RATIO WITH RFEND/ACTUAL
? 'DOING FINAL SORT BY RATIO'
SORT ON RATIO/D,BGFREQ/D,DESCRIPTOR TO FINAL
USE FINAL
*********************************************************************************
set talk off
DO CASE
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO 'Adenoid.Patent Figures:Subtraction.txt'
```

```
SET ALTERNATE ON
ENDCASE

STORE VAL(SYS(2)) TO FINTIME
IF FINTIME<STARTIME
STORE FINTIME+86400 TO FINTIME
ENDIF
STORE FINTIME - STARTIME TO COMPSEC
STORE COMPSEC/60 TO COMPMIN

*********************
SET MARGIN TO 10
@1,1 SAY 'Library Subtraction Analysis' STYLE 65536 FONT 'Geneva',274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '         '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,5,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
? Target1
IF Target2<>'
?? ', '
?? Target2
ENDIF
IF Target3<>'
?? ', '
?? Target3
ENDIF
? 'Subtracting:
? Object1
IF Object2<>'
?? ', '
?? Object2
ENDIF
IF Object3<>'
?? ', '
?? Object3
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF ANAL=1
? 'Sorted by ABUNDANCE'
ENDIF
IF ANAL=2
? 'Arranged by FUNCTION'
ENDIF
```

```
? 'Total clones represented: '
?? STR(TOT,5,0)
? 'Total clones analyzed: '
?? STR(STARTOT,5,0)
? 'Total computation time:
?? STR(COMPMIN,5,2)
?? ' minutes'
?
? 'd = designation   f = distribution   z = location   r = function   s = species   i = inte
?
************************
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',9 COLOR 0,0,0,
DO CASE
CASE ANAL=1
?? STR(AUNIQUE,4,0)
?? ' genes, for a total of '
?? STR(ANALTOT,4,0)
?? ' clones'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I
SET PRINT OFF
CLOSE DATABASES
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'

CASE ANAL=2
* arrange/function
SET PRINT ON
SET HEADING ON
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
?
? '                         BINDING PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Surface molecules and receptors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='B'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Calcium-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='C'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Ligands and effectors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='S'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'Other binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='I'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',268 COLOR 0
? '                         ONCOGENES'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'General oncogenes:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='O'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Helvetica',265 COLOR 0
? 'GTP-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='G'
```

```
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Viral elements:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='V'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Kinases and Phosphatases:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Y'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Tumor-related antigens:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='A'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',268 COLOR 0
? '           PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Transcription and Nucleic Acid-binding proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='D'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Translation:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='T'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Ribosomal proteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='R'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Protein processing:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='L'
?
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',268 COLOR 0
?
? '                    ENZYMES'
?

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Ferroproteins:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='F'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Proteases and inhibitors:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='P'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 285,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Oxidative phosphorylation:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Z'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Sugar metabolism:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 296,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='Q'

SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Helvetica',265 COLOR 0
? 'Amino acid metabolism:'
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS FONT 'Geneva',7 COLOR 0,0,0,
```

54

```
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='M'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Nucleic acid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='N'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Lipid metabolism:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='W'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other enzymes:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='E'
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",268 COLOR 0
?
? '                    MISCELLANEOUS CATEGORIES'
?

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Stress response:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='H'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Structural:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='K'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Other clones:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='X'

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Helvetica",265 COLOR 0
? 'Clones of unknown function:'
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
list OFF fields number,D,F,Z,R,ENTRY,S,DESCRIPTOR,BGFREQ,RFEND,RATIO,I FOR R='U'
ENDCASE

DO "Test print.prg"
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
ERASE TEMPLIB.DBF
ERASE TEMPNUM.DBF
ERASE TEMPDESIG.DBF
SET MARGIN TO 0
CLEAR
LOOP
ENDDO
```

```
*Northern (single), version 11-25-94
close databas s
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '             ' TO Eobject
STORE '                        ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:  8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes...: Format file Northern (single)
*
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY 'Entry #:' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,0
@ PIXELS 115,173 GET Eobject STYLE 0 FONT 'Geneva',12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY 'Description' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT 'Geneva',12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY 'Single Northern search screen' STYLE 65536 FONT 'Geneva',274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT 'Chicago',12 PICTURE '@*R Continue;Bail out' SIZE
@ PIXELS 175,98 SAY 'Clone #:' STYLE 65536 FONT 'Geneva',12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT 'Geneva',12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY 'Enter any ONE of the following:' STYLE 65536 FONT 'Geneva',12 COLOR -1;
*
* EOP: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE 'SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf'
SET TALK ON

IF Eobject<>'
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO 'Lookup entry.dbf'
SET SAFETY ON
USE 'Lookup entry.dbf'
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE 'Lookup entry.dbf'
ENDIF

IF Dobject<>'
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO 'Lookup descriptor.dbf'
SET SAFETY On
USE 'Lookup descriptor.dbf'
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
        LOOP
        ENDIF
        BROWSE
        STORE Entry TO Searchval
        CLOSE DATABASES
        ERASE 'Lookup descriptor.dbf'
        SET EXACT ON
        ENDIF

        IF Numb<>0
        USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
        GO Numb
        BROWSE
        STORE Entry TO Searchval
        ENDIF

        CLEAR
        ? 'Northern analysis for entry '
        ?? Searchval
        ?
        ? 'Enter Y to proceed'
        WAIT TO OK
        CLEAR
        IF UPPER(OK)<>'Y'
        screen 1 off
        RETURN
        ENDIF

        * COMPRESSION SUBROUTINE FOR Library.dbf
        ? 'Compressing the Libraries file now...'
        USE 'SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf'
        SET SAFETY OFF
        SORT ON library TO 'Compressed libraries.dbf'
        * FOR entered<>0
        SET SAFETY ON
        USE 'Compressed libraries.dbf'
        DELETE FOR entered=0
        PACK
        COUNT TO TOT
        MARK1 = 1
        SW2=0
        DO WHILE SW2=0 ROLL
          IF MARK1 >= TOT
          PACK
            SW2=1
            LOOP
            ENDIF
        GO MARK1
        STORE library TO TESTA
        SKIP
        STORE Library TO TESTB
        IF TESTA = TESTB
        DELETE
        ENDIF
        MARK1 = MARK1+1
        LOOP
        ENDDO ROLL

        * Northern analysis
        CLEAR
        ? 'Doing the northern now...'
        SET TALK ON
        USE 'SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf'
        SET SAFETY OFF
        COPY TO 'Hits.dbf' FOR entry=searchval
        SET SAFETY ON
```

```
* MASTER ANALYSIS 3; VERSION 12-9-94
* Master menu for analysis output
CLOSE DATABASES
SET TALK OFF
SET SAFETY OFF
CLEAR
SET DEVICE TO SCREEN
SET DEFAULT TO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:"
USE "SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf"
GO TOP
STORE NUMBER TO INITIATE
GO BOTTOM
STORE NUMBER TO TERMINATE
STORE 0 TO ENTIRE
STORE 0 TO CONDEN
STORE 0 TO ANAL
STORE 0 TO EMATCH
STORE 0 TO HMATCH
STORE 0 TO OMATCH
STORE 0 TO IMATCH
STORE 0 TO XMATCH
STORE 0 TO PRINTON
STORE 0 TO PTF
DO WHILE .T.
* Program.: Master analysis.fmt
* Date....: 12/ 9/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Master analysis
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",9 COLOR 0,0,0,
@ PIXELS 39,255 TO 277,430 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 75,120 TO 178,241 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 27,98 SAY "Customized Output Menu" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-1,-1,-1
@ PIXELS 45,54 GET conden STYLE 65536 FONT "Chicago",12 PICTURE "@*C Condensed format" SIZE
@ PIXELS 54,261 GET anal STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Sort/number;Sort/entry;
@ PIXELS 117,126 GET EMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Exact " SIZE 15,62 CO
@ PIXELS 135,126 GET HMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Homologous" SIZE 15,1
@ PIXELS 153,126 GET OMATCH STYLE 65536 FONT "Chicago",12 PICTURE "@*C Other spc" SIZE 15,84
@ PIXELS 90,152 SAY "Matches:" STYLE 65536 FONT "Geneva",268 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 63,54 GET PRINTON STYLE 65536 FONT "Chicago",12 PICTURE "@*C Include clone listing"
@ PIXELS 171,126 GET Imatch STYLE 65536 FONT "Chicago",12 PICTURE "@*C Incyte" SIZE 15,65 CO
@ PIXELS 252,146 GET initiate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,146 GET terminate STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 234,134 SAY "Include clones " STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 270,125 SAY "-->" STYLE 65536 FONT "Geneva",14 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 198,126 GET PTF STYLE 65536 FONT "Chicago",12 PICTURE "@*C Print to file" SIZE 15,9
@ PIXELS 189,0 TO 257,120 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 209,8 SAY "Library selection" STYLE 65536 FONT "Geneva",266 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 227,18 GET ENTIRE STYLE 65536 FONT "Chicago",12 PICTURE "@*RV All;Selected" SIZE 16
*
* EOF: Master analysis.fmt
READ
  IF ANAL=9
  CLEAR
  CLOSE DATABASES
  ERASE TEMPMASTER.DBF
  USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
  SET SAFETY ON
  SCREEN 1 OFF
  RETURN
  ENDIF
clear
? INITIATE
? TERMINATE
? CONDEN
? ANAL
```

5 8

```
? ematch
? Hmatch
? Omatch
? IMATCH
SET TALK ON
  IF ENTIRE=2
USE 'Unique libraries.dbf'
  REPLACE ALL i WITH ' '
  BROWSE FIELDS i,libname,library,total,entered AT 0,0
  ENDIF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
*COPY TO TEMPNUM FOR NUMBER>=INITIATE.AND.NUMBER<=TERMINATE
*USE TEMPNUM
COPY STRUCTURE TO TEMPLIB
USE TEMPLIB
  IF ENTIRE=1
  APPEND FROM 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf'
  ENDIF
  IF ENTIRE=2
USE 'Unique libraries.dbf'
  COPY TO SELECTED FOR UPPER(i)='Y'
  USE SELECTED
  STORE RECCOUNT() TO STOPIT
  MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    CLEAR
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    STORE library TO THISONE
    ? 'COPYING '
    ?? THISONE
    USE TEMPLIB
    APPEND FROM 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf' FOR library=THISONE
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
  ENDIF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
COUNT TO STARTOT
COPY STRUCTURE TO TEMPDESIG
USE TEMPDESIG
  IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
  APPEND FROM TEMPLIB
  ENDIF
  IF Ematch=1
  APPEND FROM TEMPLIB FOR D='E'
  ENDIF
  IF Hmatch=1
  APPEND FROM TEMPLIB FOR D='H'
  ENDIF
  IF Omatch=1
  APPEND FROM TEMPLIB FOR D='O'
  ENDIF
  IF Imatch=1
  APPEND FROM TEMPLIB FOR D='I'.OR.D='X'.OR.D='N'
  ENDIF
  IF Xmatch=1
  APPEND FROM TEMPLIB FOR D='X'
  ENDIF
COUNT TO ANALTOT
set talk off
********************
DO CASE
```

5 9

```
CASE PTF=0
SET DEVICE TO PRINT
SET PRINT ON
EJECT
CASE PTF=1
SET ALTERNATE TO "Total function sort.txt"
*SET ALTERNATE TO "H and O function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Abundance con.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Function sort.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Distribution sort.txt"
*SET ALTERNATE TO "Shear stress HUVEC 1:Clone list.txt"
*SET ALTERNATE TO "Shear Stress HUVEC 2:Location sort.txt"
SET ALTERNATE ON
ENDCASE
*********************
IF PRINTON=1
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
ENDIF
?
?
?
?
? date()
?? '   '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
    MARK=1
    DO WHILE .T.
    IF MARK>STOPIT
    EXIT
    ENDIF
    USE SELECTED
    GO MARK
    ? '   '
    ?? TRIM(libname)
    STORE MARK+1 TO MARK
    LOOP
    ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0 .AND. IMATCH=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF Imatch=1
?? 'INCYTE'
ENDIF
IF Xmatch=1
?? 'EST'
```

6 0

```
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
? '1 = library    d = designation   f = distribution    z = location    r = function    c = cer
?
***************************************************************************************************
USE TEMPDESIG
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
DO CASE
CASE ANAL=1
* sort/number
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER
DO "COMPRESSION number.PRG"
ELSE
SORT TO TEMP1 ON NUMBER
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR
*list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=2
* sort/DESCRIPTOR
SET HEADING ON
*SORT TO TEMP1 ON DESCRIPTOR,ENTRY,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
*SORT TO TEMP1 ON ENTRY,DESCRIPTOR,NUMBER/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
SORT TO TEMP1 ON ENTRY,START/S for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
IF CONDEN=1
DO "COMPRESSION entry.PRG"
ELSE
USE TEMP1
 list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF
```

6 1

```
CASE ANAL=3
* sort by abundance
SET HEADING ON
SORT TO TEMP1 ON ENTRY,NUMBER for D='E'.OR.D='H'.OR.D='O'.OR.D='X'.OR.D='I'
DO "COMPRESSION abundance.PRG"

CASE ANAL=4.
* sort/interest
SET HEADING ON
IF CONDEN=1
SORT TO TEMP1 ON ENTRY,NUMBER FOR I>0
DO "COMPRESSION interest.PRG"
ELSE
SORT ON I/D,ENTRY TO TEMP1 FOR I>1
USE TEMP1
list off fields number,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,RFEND,INIT,I
CLOSE DATABASES
ERASE TEMP1.DBF
ENDIF

CASE ANAL=5
* arrange/location
SET HEADING ON
STORE 4 TO AMPLIFIER
? 'Nuclear:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR.LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoplasmic:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cytoskeleton:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Cell surface:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Intracellular membrane:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Mitochondrial:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
```

```
? 'Secreted;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression location.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
EJECT
DO "Output heading.prg"
USE "Analysis location.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '      FUNCTIONAL CLASS                      TOTAL   UNIQUE  NEW  % TOTAL'
?
LIST OFF FIELDS Z,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF

CASE ANAL=6
* arrange/distribution
SET HEADING ON
STORE 3 TO AMPLIFIER
? 'Cell/tissue specific distribution;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Non-specific distribution;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Unknown distribution;'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression distrib.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
IF CONDEN=1
SET DEVICE TO PRINTER
SET PRINTER ON
```

```
EJECT
DO "Output heading.prg"
USE "Analysis distribution.dbf"
DO "Create bargraph.prg"
SET HEADING OFF
? '          FUNCTIONAL CLASS                    TOTAL   UNIQUE  % TOTAL'
?
LIST OFF FIELDS P.NAME,CLONES,GENES,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF


CASE ANAL=7
* arrange/function
SET HEADING ON
STORE 10 TO AMPLIFIER
? '                         BINDING PROTEINS'
?
? 'Surface molecules and receptors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Calcium-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ligands and effectors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                         ONCOGENES'
?
? 'General oncogenes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'GTP-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Viral elements:'
```

```
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Kinases and Phosphatases:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Tumor-related antigens:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              PROTEIN SYNTHETIC MACHINERY PROTEINS'
?
? 'Transcription and Nucleic Acid-binding proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Translation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Ribosomal proteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Protein processing:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg".
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              ENZYMES'
?
? 'Ferroproteins:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Proteases and inhibitors:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compressi n function.prg"
ELSE
```

```
DO "Normal subroutine 1"
ENDIF
? 'Oxidative phosphorylation:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Sugar metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Amino acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Nucleic acid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Lipid metabolism:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other enzymes:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
*EJECT
? '                              MISCELLANEOUS CATEGORIES'
?
? 'Stress response:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Structural:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression function.prg"
ELSE
DO "Normal subroutine 1"
ENDIF
? 'Other clones:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO "Compression functi n.prg"
ELSE
```

66

```
DO 'Normal subroutine 1'
ENDIF
? 'Clones of unknown function:'
SORT ON ENTRY,NUMBER FIELDS RFEND,NUMBER,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I,COMMEN
IF CONDEN=1
DO 'Compression function.prg'
ELSE
DO 'Normal subroutine 1'
ENDIF

IF CONDEN=1
EJECT
*SET DEVICE TO PRINTER
*SET PRINT ON
DO 'Output heading.prg'
***
USE 'Analysis function.dbf'
DO 'Create bargraph.prg'
SET HEADING OFF
***
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
***
? '                                                     TOTAL    TOTAL    NEW      DIST
? '       FUNCTIONAL CLASS                       CLONES  GENES   GENES   FUNCTIONAL CLASS'
? '
***
*LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH,COMPANY
LIST OFF FIELDS P,NAME,CLONES,GENES,NEW,PERCENT,GRAPH
CLOSE DATABASES
ERASE TEMP2.DBF
SET HEADING ON
*USE "SmartGuy:FoxBASE+/Mac:fox files:TEMPMASTER.dbf"
ENDIF
CASE ANAL=8
DO 'Subgroup summary 3.prg'
ENDCASE
DO 'Test print.prg'
SET PRINT OFF
SET DEVICE TO SCREEN
CLOSE DATABASES
*ERASE TEMPLIB.DBF
*ERASE TEMPNUM.DBF
*ERASE TEMPDESIG.DBF
*ERASE SELECTED.DBF
CLEAR
LOOP
ENDDO
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   COUNT TO NEWGENES FOR D='H'.OR.D='O'
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1.
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE Z TO LOC
USE 'Analysis location.dbf'
LOCATE FOR Z=LOC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' .clones'
? '                          V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON DATE TO TEMP2
USE TEMP2
?? STR(UNIQUE,4,0)
?? ' genes, for a total of'
?? STR(TOT,4,0)
?? ' clones'
?
? '                    V Coincidence'
COUNT TO P4 FOR I=4
IF P4>0
? STR(P4,3,0)
?? ' genes with priority = 4 (Secondary analysis:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=4
?
ENDIF
COUNT TO P3 FOR I=3
IF P3>0
? STR(P3,3,0)
?? ' genes with priority = 3 (Full insert sequence:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=3
?
ENDIF
COUNT TO P2 FOR I=2
IF P2>0
? STR(P2,3,0)
?? ' genes with priority = 2 (Primary analysis complete:)'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=2
?
ENDIF
COUNT TO P1 FOR I=1
IF P1>0
```

```
? STR(P1,3,0)
?? ' genes with priority = 1 (Primary analysis needed:)'
list off fields number,RFEND,L,D,P,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT for I=1
ENDIF


*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf*
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON NUMBER TO TEMP2
USE TEMP2

?? STR(UNIQUE,4,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                     V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   COUNT TO NEWGENES FOR D='H'.OR.D='O'
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE R TO FUNC
USE "Analysis function.dbf"
LOCATE FOR P=FUNC
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
REPLACE NEW WITH NEWGENES.
USE TEMP1
SORT ON RFEND/D TO TEMP2
USE TEMP2
SET HEADING ON
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
***
? '                           V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I
***

*SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,
*list off fields RFEND,S,DESCRIPTOR

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

72

```
* CCMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
STORE F TO DIST
USE "Analysis distribution.dbf"
LOCATE FOR P=DIST
REPLACE CLONES WITH TOT
REPLACE GENES WITH UNIQUE
USE TEMP1
sort on rfend/d to TEMP2
USE TEMP2
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '                        V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
   ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
GO TOP
USE TEMP1
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
?  '                        V Coincidence'
list off fields number,RFEND,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,LENGTH,INIT,I

*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE 'SmartGuy:FoxBASE+/Mac:fox files:Clones.dbf'
COPY TO TEMP1 FOR
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   COUNT TO UNIQUE
   SW2=1
   LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
   IF TESTA = TESTB
   DELETE
   DUP = DUP+1
   LOOP
   ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? ' Coincidence V        V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR,INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

```
* COMPRESSION SUBROUTINE FOR ANALYSIS PROGRAMS
USE TEMP1
COUNT TO IDGENE FOR D='E'.OR.D='O'.OR.D='H'.OR.D='N'.OR.D='R'.OR.D='A'
DELETE FOR D='N'.OR.D='D'.OR.D='A'.OR.D='U'.OR.D='S'.OR.D='M'.OR.D='R'.OR.D='V'
PACK
COUNT TO TOT
REPLACE ALL RFEND WITH 1
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
  IF MARK1 >= TOT
  PACK
  COUNT TO UNIQUE
  SW2=1
  LOOP
  ENDIF
GO MARK1
DUP = 1
STORE ENTRY TO TESTA
SW = 0
DO WHILE SW=0 TEST
SKIP
STORE ENTRY TO TESTB
  IF TESTA = TESTB
  DELETE
  DUP = DUP+1
  LOOP
  ENDIF
GO MARK1
REPLACE RFEND WITH DUP
MARK1 = MARK1+DUP
SW=1
LOOP
ENDDO TEST
LOOP
ENDDO ROLL
*BROWSE
*SET PRINTER ON
SORT ON RFEND/D,NUMBER TO TEMP2
USE TEMP2
REPLACE ALL START WITH RFEND/IDGENE*10000
?? STR(UNIQUE,5,0)
?? ' genes, for a total of '
?? STR(TOT,5,0)
?? ' clones'
? '  Coincidence V          V Clones/10000'
set heading off
SCREEN 1 TYPE 0 HEADING 'Screen 1' AT 40,2 SIZE 286,492 PIXELS  FONT 'Geneva',7 COLOR 0,0,0,
list fields number,RFEND,START,L,D,F,Z,R,C,ENTRY,S,DESCRIPTOR;INIT,I
*SET PRINT OFF
CLOSE DATABASES
ERASE TEMP1.DBF
ERASE TEMP2.DBF
USE 'SmartGuy:FoxBASE+/Mac:fox files:clones.dbf'
```

7 6

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE.TEMP1.DBF
USE TEMPDESIG
```

```
*Lifescan menu; version 8-7-94
SET TALK OFF
set device to screen
CLEAR
USE "SmartGuy:FoxBASE+/Mac:fox files:clones.dbf"
STORE LUPDATE() TO Update
GO BOTTOM
STORE RECNO() TO cloneno
STORE 6 TO Chooser
DO WHILE .T.
* Program.: Lifeseq menu.fmt
* Date....:  1/11/95
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Lifeseq menu
*
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",268 COLOR 0,0,
@ PIXELS 18,126 TO 77,365 STYLE 28479 COLOR 32767,-25600,-1,-16223,-16721,-15725
@ PIXELS 110,29 TO 188,217 STYLE 3871 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 45,161 SAY "LIFESEQ" STYLE 65536 FONT "Geneva",536 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 36,269 SAY "TM" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,7135,5884
@ PIXELS 63,143 SAY "Molecular Biology Desktop" STYLE 65536 FONT "Helvetica",18 COLOR 0,0,0,
@ PIXELS 90,252 TO 251,467 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 117,270 GET Chooser STYLE 65536 FONT "Chicago",12 PICTURE "@*RV Transcript profiles
@ PIXELS 135,128 SAY Update STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 171,128 SAY cloneno STYLE 0 FONT "Geneva",12 SIZE 15,79 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 135,44 SAY "Last update:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 171,44 SAY "Total clones:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,-1,-1,-1,-1
@ PIXELS 45,296 SAY "v1.30" STYLE 65536 FONT "Geneva",782 COLOR 0,0,-1,-1,-1,-1
*
* EOF: Lifeseq menu.fmt
READ
DO CASE
CASE Chooser=1
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Master analysis 3.prg"
CASE Chooser=2
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Subtraction 2.prg"
CASE Chooser=3
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:Northern (single).prg"
CASE Chooser=4
USE "Libraries.dbf"
BROWSE
CASE Chooser=5
DO "SmartGuy:FoxBASE+/Mac:fox files:Output programs:See individual clone.prg"
CASE Chooser=6
DO "SmartGuy:FoxBASE+/Mac:fox files:Libraries:Output programs:Menu.prg"
CASE Chooser=7
CLEAR
SCREEN 1 OFF
RETURN
ENDCASE

LOOP
ENDDO
```

7 8

```
@1,30 SAY "Database Subset Analysis" STYLE 65536 FONT "Geneva",274 COLOR 0,0,0,-1,-1,-1
?
?
?
?
? date()
?? '        '
?? TIME()
? 'Clone numbers '
?? STR(INITIATE,6,0)
?? ' through '
?? STR(TERMINATE,6,0)
? 'Libraries: '
IF ENTIRE=1
? 'All libraries'
ENDIF
IF ENTIRE=2
     MARK=1
     DO WHILE .T.
     IF MARK>STOPIT
     EXIT
     ENDIF
     USE SELECTED
     GO MARK
     ? '    '
     ?? TRIM(libname)
     STORE MARK+1 TO MARK
     LOOP
     ENDDO
ENDIF
? 'Designations: '
IF Ematch=0 .AND. Hmatch=0 .AND. Omatch=0
?? 'All'
ENDIF
IF Ematch=1
?? 'Exact,'
ENDIF
IF Hmatch=1
?? 'Human,'
ENDIF
IF Omatch=1
?? 'Other sp.'
ENDIF
IF CONDEN=1
? 'Condensed format analysis'
ENDIF
IF ANAL=1
? 'Sorted by NUMBER'
ENDIF
IF ANAL=2
? 'Sorted by ENTRY'
ENDIF
IF ANAL=3
? 'Arranged by ABUNDANCE'
ENDIF
IF ANAL=4
? 'Sorted by INTEREST'
ENDIF
IF ANAL=5
? 'Arranged by LOCATION'
ENDIF
IF ANAL=6
? 'Arranged by DISTRIBUTION'
ENDIF
IF ANAL=7
? 'Arranged by FUNCTION'
```

```
ENDIF
? 'Total clones represented: '
?? STR(STARTOT,6,0)
? 'Total clones analyzed: '
?? STR(ANALTOT,6,0)
?
?
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones'
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
USE TEMP1
COUNT TO TOT
?? ' Total of'
?? STR(TOT,4,0)
?? ' clones!
?
*list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR,LENGTH,RFEND,INIT,I
list off fields number,L,D,F,Z,R,C,ENTRY,DESCRIPTOR
CLOSE DATABASES
ERASE TEMP1.DBF
USE TEMPDESIG
```

```
*Northern (single), version 11-25-94
close databases
SET TALK OFF
SET PRINT OFF
SET EXACT OFF
CLEAR
STORE '           ' TO Eobject
STORE '                        ' TO Dobject
STORE 0 TO Numb
STORE 0 TO Zog
STORE 1 TO Bail
DO WHILE .T.
* Program.: Northern (single).fmt
* Date....:   8/ 8/94
* Version.: FoxBASE+/Mac, revision 1.10
* Notes....: Format file Northern (single)
*

SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",12 COLOR 0,0,0
@ PIXELS 15,81 TO 46,397 STYLE 28447 COLOR 0,0,-1,-25600,-1,-1
@ PIXELS 89,79 TO 192,422 STYLE 28447 COLOR 0,0,0,-25600,-1,-1
@ PIXELS 115,98 SAY "Entry #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 115,173 GET Eobject STYLE 0 FONT "Geneva",12 SIZE 15,142 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,89 SAY "Description" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 145,173 GET Dobject STYLE 0 FONT "Geneva",12 SIZE 15,241 COLOR 0,0,0,-1,-1,-1
@ PIXELS 35,89 SAY "Single Northern search screen" STYLE 65536 FONT "Geneva",274 COLOR 0,0,-
@ PIXELS 220,162 GET Bail STYLE 65536 FONT "Chicago",12 PICTURE "@*R Continue;Bail out" SIZE
@ PIXELS 175,98 SAY "Clone #:" STYLE 65536 FONT "Geneva",12 COLOR 0,0,0,-1,-1,-1
@ PIXELS 175,173 GET Numb STYLE 0 FONT "Geneva",12 SIZE 15,70 COLOR 0,0,0,-1,-1,-1
@ PIXELS 80,152 SAY "Enter any ONE of the following:" STYLE 65536 FONT "Geneva",12 COLOR -1,

* EOF: Northern (single).fmt
READ
IF Bail=2
CLEAR
screen 1 off
RETURN
ENDIF
USE "SmartGuy:FoxBASE+/Mac:Fox files:Lookup.dbf"
SET TALK ON

IF Eobject<>'          '
STORE UPPER(Eobject) to Eobject
SET SAFETY OFF
SORT ON Entry TO "Lookup entry.dbf"
SET SAFETY ON
USE "Lookup entry.dbf"
LOCATE FOR Look=Eobject
IF .NOT.FOUND()
CLEAR
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup entry.dbf"
ENDIF

IF Dobject<>'                        '
SET EXACT OFF
SET SAFETY OFF
SORT ON descriptor TO "Lookup descriptor.dbf"
SET SAFETY On
USE "Lookup descriptor.dbf"
LOCATE FOR UPPER(TRIM(descriptor))=UPPER(TRIM(Dobject))
IF .NOT.FOUND()
CLEAR
```

```
LOOP
ENDIF
BROWSE
STORE Entry TO Searchval
CLOSE DATABASES
ERASE "Lookup descriptor.dbf"
SET EXACT ON
ENDIF

IF Numb<>0
USE "SmartGuy:FoxBASE+/Mac:Fox files:clones.dbf"
GO Numb
BROWSE
STORE Entry TO Searchval
ENDIF

CLEAR
? 'Northern analysis for entry '
?? Searchval
?
? 'Enter Y to proceed'
WAIT TO OK
CLEAR
IF UPPER(OK)<>'Y'
screen 1 off
RETURN
ENDIF

* COMPRESSION SUBROUTINE FOR Library.dbf
? 'Compressing the Libraries file now...'
USE "SmartGuy:FoxBASE+/Mac:Fox files:libraries.dbf"
SET SAFETY OFF
SORT ON library TO "Compressed libraries.dbf"
* FOR entered>0
SET SAFETY ON
USE "Compressed libraries.dbf"
DELETE FOR entered=0
PACK
COUNT TO TOT'
MARK1 = 1
SW2=0
DO WHILE SW2=0 ROLL
   IF MARK1 >= TOT
   PACK
   SW2=1
   LOOP
   ENDIF
GO MARK1
STORE library TO TESTA
SKIP
STORE Library TO TESTB
IF TESTA = TESTB
DELETE
ENDIF
MARK1 = MARK1+1
LOOP
ENDDO ROLL

* Northern analysis
CLEAR
? 'Doing the northern now...'
SET TALK ON
USE "SmartGuy:FoxBASE+/Mac:F x files:cl nes.dbf"
SET SAFETY OFF
COPY TO "Hits.dbf" FOR entry=searchval
SET SAFETY ON
```

84

```
CLOSE DATABASES
SELECT 1
USE "Compressed libraries.dbf"
STORE RECCOUNT() TO Entries
SELECT 2
USE "Hits.dbf"
Mark=1
DO WHILE .T.
SELECT 1
IF Mark>Entries
EXIT
ENDIF
GO MARK
STORE library TO Jigger
SELECT 2
COUNT TO Zog FOR library=Jigger
SELECT 1
REPLACE hits with Zog
Mark=Mark+1
LOOP
ENDDO

SELECT 1
BROWSE FIELDS LIBRARY,LIBNAME,ENTERED,HITS AT 0,0
CLEAR
? 'Enter Y to print:'
WAIT TO PRINSET
IF UPPER(PRINSET)='Y'
SET PRINT ON
CLEAR
EJECT
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",14 COLOR 0,0,0
? 'DATABASE ENTRIES MATCHING ENTRY '
?? Searchval
? DATE()
?
SCREEN 1 TYPE 0 HEADING "Screen 1" AT 40,2 SIZE 286,492 PIXELS  FONT "Geneva",7 COLOR 0,0,0,
LIST OFF FIELDS library,libname,entered,hits
?
?
SELECT 2
LIST OFF FIELDS NUMBER,LIBRARY,D,S,F,Z,R,ENTRY,DESCRIPTOR,RFSTART,START,RFEND
SET TALK OFF
SET PRINT OFF
ENDIF
CLOSE DATABASES
SET TALK OFF
CLEAR
DO "Test print.prg"
RETURN
```

85

TABLE 6

library	libname
ADENINB01	Inflamed adenoid
ADRENOR01	Adrenal gland (r)
ADRENOT01	Adrenal gland (T)
AMLBNOT01	AML blast cells (T)
BMARNOT01	Bone marrow
BMARNOT02	Bone marrow (T)
CARDNOT01	Cardiac muscle (T)
CHAONOT01	Chin. hamster ovary
CORNNOT01	Corneal stroma
FIBRAGT01	Fibroblast, AT 5
FIBRAGT02	Fibroblast, AT 30
FIBRANT01	Fibroblast, AT
FIBRNGT01	Fibroblast, uv 5
FIBRNGT02	Fibroblast, uv 30
FIBRNOT01	Fibroblast
FIBRNOT02	Fibroblast, normal
HMC1NOT01	Mast cell line HMC-1
HUVELPB01	HUVEC IFN,TNF,LPS
HUVENOB01	HUVEC control
HUVESTB01	HUVEC shear stress
HYPONOB01	Hypothalamus
KIDNNOT01	Kidney (T)
LIVRNOT01	Liver (T)
LUNGNOT01	Lung (T)
MUSCNOT01	Skeletal muscle (T)
OVIDNOB01	Oviduct
PANCNOT01	Pancreas, normal
PITUNOR01	Pituitary (r)
PITUNOT01	Pituitary (T)
PLACNOB01	Placenta
SINTNOT02	Small intestine (T)
SPLNFET01	Spleen+liver, fetal
SPLNNOT02	Spleen (T)
STOMNOT01	Stomach
SYNORAB01	Rheum. synovium
TBLYNOT01	T + B lymphoblast
TESTNOT01	Testis (T)
THP1NOB01	THP-1 control
THP1PEB01	THP phorbol
THP1PLB01	THP-1 phorbol LPS
U937NOT01	U937, monocytic leuk

number	library	d	s	f	z	r	entry	descriptor	rfstart	fstart	rfend
2304	U937NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	0	773
3240	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	370	773
3269	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	371	773
4693	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	470	773
8989	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	327	773
9139	HMC1NOT01	E	H	C	C	T	HUMEF1B	Elongation factor 1-beta	0	375	773

WHAT IS CLAIMED IS:

1.   A method of analyzing a specimen containing gene transcripts, said method comprising the steps of:

    (a) producing a library of biological sequences;

    (b) generating a set of transcript sequences, where each of the transcript sequences in said set is indicative of a different one of the biological sequences of the library;

    (c) processing the transcript sequences in a programmed computer in which a database of reference transcript sequences indicative of reference biological sequences is stored, to generate an identified sequence value for each of the transcript sequences, where each said identified sequence value is indicative of a sequence annotation and a degree of match between one of the transcript sequences and at least one of the reference transcript sequences; and

    (d) processing each said identified sequence value to generate final data values indicative of a number of times each identified sequence value is present in the library.

2.   The method of claim 1, wherein step (a) includes the steps of:

    obtaining a mixture of mRNA;

    making cDNA copies of the mRNA;

    isolating a representative population of clones transfected with the cDNA and producing therefrom the library of biological sequences.

3.   The method of claim 1, wherein the biological sequences are cDNA sequences.

4.   The method of claim 1, wherein the biological sequences are RNA sequences.

5.   The method of claim 1, wherein the biological sequences are protein sequences.

87

6.   The method of claim 1, wherein a first value of said degree of match is indicative of an exact match, and a second value of said degree of match is indicative of a non-exact match.

5   7.   A method of comparing two specimens containing gene transcripts, said method comprising:
        (a)   analyzing a first specimen according to the method of claim 1;
        (b)   producing a second library of biological
10 sequences;
        (c)   generating a second set of transcript sequences, where each of the transcript sequences in said second set is indicative of a different one of the biological sequences of the second library;
15 ·   (d)   processing the second set of transcript sequences in said programmed computer to generate a second set of identified sequence values known as further identified sequence values, where each of the further identified sequence values is indicative of a sequence annotation and
20 a degree of match between one of the biological sequences of the second library and at least one of the reference sequences;
        (e)   processing each said further identified sequence value to generate further final data values indicative of a
25 number of times each further identified sequence value is present in the second library; and
        (f)   processing the final data values from the first specimen and the further identified sequence values from the second specimen to generate ratios of transcript
30 sequences, each of said ratio values indicative of differences in numbers of gene transcripts between the two specimens.

        8.   A method of quantifying relative abundance of mRNA in a biological specimen, said method comprising the steps
35 of:
        (a)   isolating a population of mRNA transcripts from the biological specimen;

(b)   identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c)   determining numbers of mRNA transcripts corresponding to each of the genes; and

5   (d)   using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts.

9.   A diagnostic method which comprises producing a gene transcript image, said method comprising the steps of:

10   (a)   isolating a population of mRNA transcripts from a biological specimen;

(b)   identifying genes from which the mRNA was transcribed by a sequence-specific method;

(c)   determining numbers of mRNA transcripts
15 corresponding to each of the genes; and

(d)   using the mRNA transcript numbers to determine the relative abundance of mRNA transcripts within the population of mRNA transcripts, where data determining the relative abundance values of mRNA transcripts is the gene
20 transcript image of the biological specimen.

10.   The method of claim 9, further comprising:

(e)   providing a set of standard normal and diseased gene transcript images; and

(f)   comparing the gene transcript image of the
25 biological specimen with the gene transcript images of step (e) to identify at least one of the standard gene transcript images which most closely approximate the gene transcript image of the biological specimen.

11.   The method of claim 9, wherein the biological
30 specimen is biopsy tissue, sputum, blood or urine.

12.   A method of producing a gene transcript image, said method comprising the steps of

(a)   obtaining a mixture of mRNA;

(b)   making cDNA copies of the mRNA;

(c)   inserting the cDNA into a suitable vector and using said vector to transfect suitable host strain cells which are plated out and permitted to grow into clones, each clone representing a unique mRNA;

5   (d)   isolating a representative population of recombinant clones;

(e)   identifying amplified cDNAs from each clone in the population by a sequence-specific method which identifies gene from which the unique mRNA was transcribed;

10   (f)   determining a number of times each gene is represented within the population of clones as an indication of relative abundance; and

(g)   listing the genes and their relative abundance in order of abundance, thereby producing the gene transcript
15   image.

13.   The method of claim 12, also including the step of diagnosing disease by:

repeating steps (a) through (g) on biological specimens from random sample of normal and diseased humans,
20   encompassing a variety of diseases, to produce reference sets of normal and diseased gene transcript images;

obtaining a test specimen from a human, and producing a test gene transcript image by performing steps (a) through (g) on said test specimen;

25   comparing the test gene transcript image with the reference sets of gene transcript images; and

identifying at least one of the reference gene transcript images which most closely approximates the test gene transcript image.

30   14.   A computer system for analyzing a library of biological sequences, said system including:

means for receiving a set of transcript sequences, where each of the transcript sequences is indicative of a different one of the biological sequences of the library;
35   and

means for processing the transcript sequences in the computer system in which a database of reference transcript

sequences indicative of reference biological sequences is
stored, wherein the computer is programmed with software
for generating an identified sequence value for each of the
transcript sequences, where each said identified sequence
5   value is indicative of a sequence annotation and a degree
of match between a different one of the biological
sequences of the library and at least one of the reference
transcript sequences, and for processing each said
identified sequence value to generate final data values
10   indicative of a number of times each identified sequence
value is present in the library.

15.   The system of claim 14, also including:
library generation means for producing the library of
biological sequences and generating said set of transcript
15   sequences from said library.

16.   The system of claim 15, wherein the library
generation means includes:
means for obtaining a mixture of mRNA;
means for making cDNA copies of the mRNA;
20   means for inserting the cDNA copies into cells and
permitting the cells to grow into clones;
means for isolating a representative population of the
clones and producing therefrom the library of biological
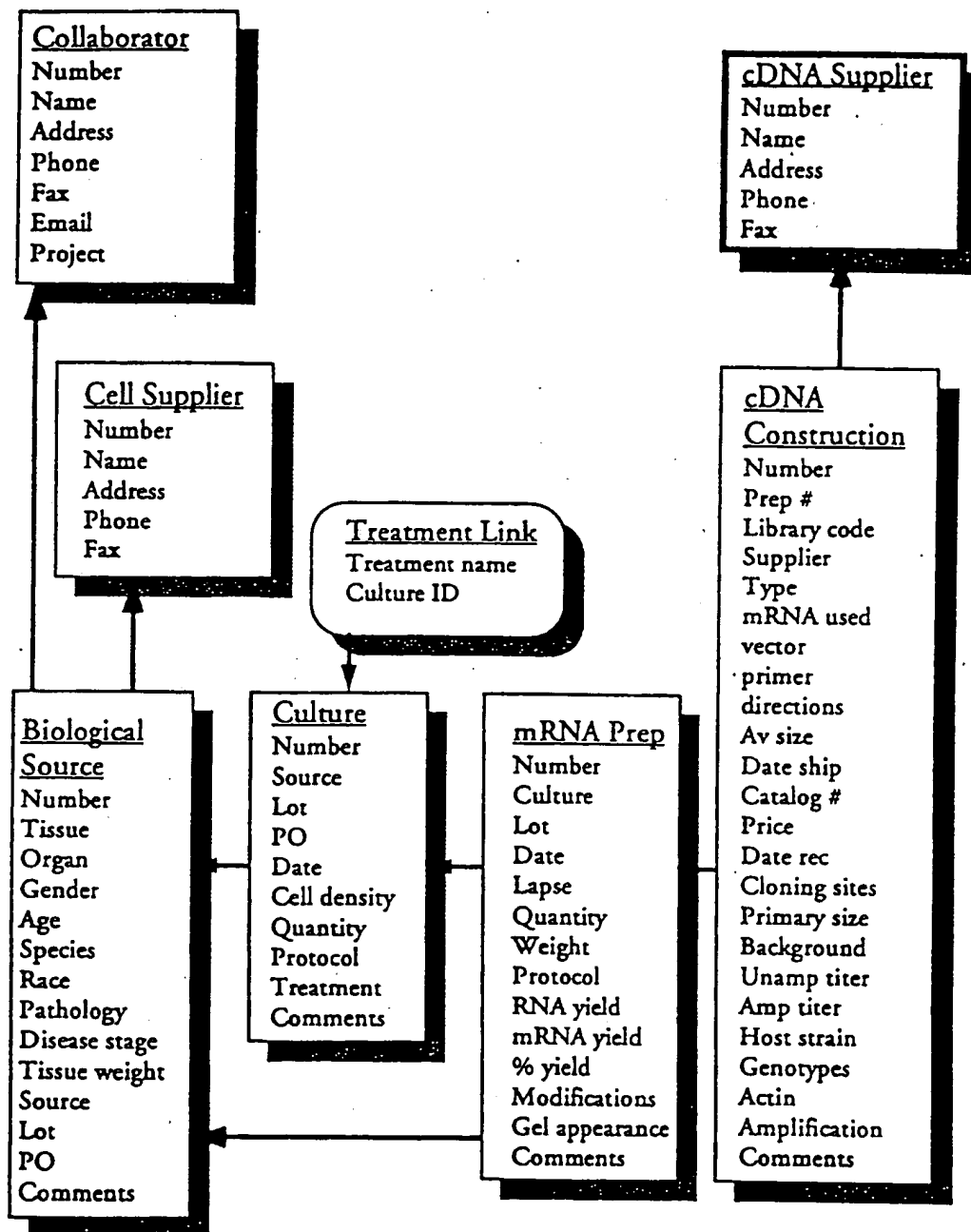sequences.
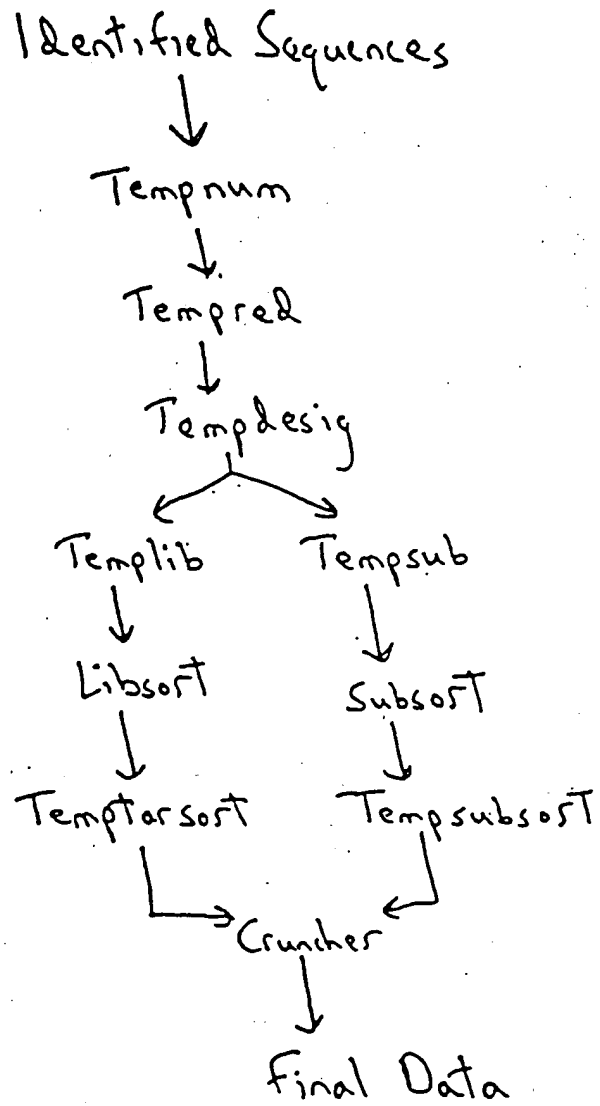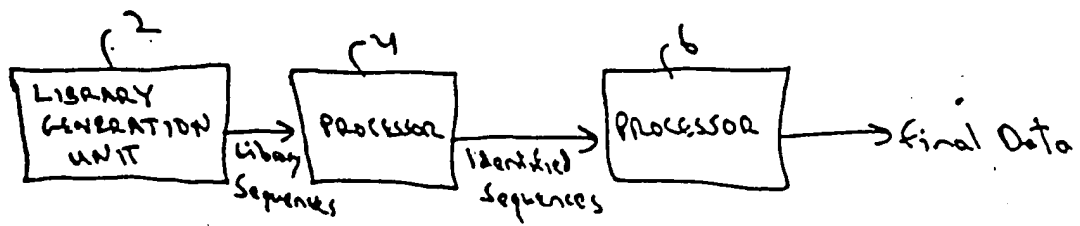
SYBASE database Structure
Library Preparation

Collaborator
Number
Name
Address
Phone
Fax
Email
Project

cDNA Supplier
Number
Name
Address
Phone
Fax

Cell Supplier
Number
Name
Address
Phone
Fax

Treatment Link
Treatment name
Culture ID

cDNA Construction
Number
Prep #
Library code
Supplier
Type
mRNA used
vector
primer
directions
Av size
Date ship
Catalog #
Price
Date rec
Cloning sites
Primary size
Background
Unamp titer
Amp titer
Host strain
Genotypes
Actin
Amplification
Comments

Biological Source
Number
Tissue
Organ
Gender
Age
Species
Race
Pathology
Disease stage
Tissue weight
Source
Lot
PO
Comments

Culture
Number
Source
Lot
PO
Date
Cell density
Quantity
Protocol
Treatment
Comments

mRNA Prep
Number
Culture
Lot
Date
Lapse
Quantity
Weight
Protocol
RNA yield
mRNA yield
% yield
Modifications
Gel appearance
Comments

Figure 1

1 / 4

Identified Sequences

↓

Tempnum

↓

Tempred

↓

Tempdesig

Templib   Tempsub

↓   ↓

Libsort   Subsort

↓   ↓

Temptarsort   Tempsubsort

Cruncher

↓

Final Data

Figure 2

Figure 3

Incyte Bioinformatics Process

Figure 4

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) :C12Q 1/68; G06F 15/00
US CL : 435/6; 364/413.02

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 364/413.02

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CAS ONLINE, APS, transcript, transcripts, cdan#, mrna#, frequenc?, distribut?, abundanc?

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	IntelliGenetics Suite, Release 5.4, Advanced Training Manual, issued January 1993 by IntelliGenetics, Inc. 700 East El Camino Real, Mountain View, California 94040, United States of America, pages (1-6)-(1-19) and (2-9)-(2-14), see entire document.	15 and 16 ---------- 1-14
Y	Science, Volume 252, issued 21 June 1991, M.D. Adams et al, "Complementary DNA sequencing: Expressed sequence tags and human genome project", pages 1651-1656, see entire document.	1-16

☒ Further documents are listed in the continuation of Box C.   ☐   See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 APRIL 1995	04 MAY 1995

Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	Authorized officer JAMES MARTINELL
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Nucleic Acids Research, Volume 19, No. 25, issued 1991, E. Hara et al, "Subtractive cDNA cloning using oligo(dT)$_{30}$-latex and PCR: isolation of cDNA clones specific to undifferentiated human embryonal carcinoma cells", pages 7097-7104, see entire document.	1-16
X — Y	Nature Genetics, Volume 2, No. 3, issued November 1992, K. Okubo et al, "Large scale cDNA sequencing for analysis of quantitative and qualitative aspects of gene expression", pages 173-179, see narrative text portion of entire document.	1, 3 ---------- 2 and 4-16

REPORTS

Axl1p sequence following Ser²⁰⁹ and occurs within the domain of Axl1p that shows homology with hIDE (14). To delete the complete STE23 sequence and create the sta23Δ::URA3 mutation, polymerase chain reaction (PCR) primers (5'-TCGGAAGACCTCAT-TCTTGCTCATTTTGATATTGCTC- TGTAGATTG-TACTGAGAGTGCAC-3'; and 5'-GCTACAAACAGC-GTCGACTTGAATGCCCCGACATCTTCGACTGT-GCGGTATTTCACACCG-3') were used to amplify the URA3 sequence of pRS316, and the reaction product was transformed into yeast for one-step gene replacement [R. Rothstein, Methods Enzymol. 194, 281 (1991)]. To create the axl1Δ::LEU2 mutation contained on p114, a 5.0-kb Sal I fragment from pAXL1 was cloned into pUC19, and an internal 4.0-kb Hpa I–Xho I fragment was replaced with a LEU2 fragment. To construct the sta23Δ::LEU2 allele (a deletion corresponding to 931 amino acids) carried on p153, a LEU2 fragment was used to replace the 2.8-kb Pml I–Ed136 II fragment of STE23, which occurs within a 6.2-kb Hind III–Bgl II genomic fragment carried on pSP72 (Promega). To create YEpMFA1, a 1.6-kb Bam HI fragment containing MFA1, from pKK16 [K. Kuchler, R. E. Sterne, J. Thorner, EMBO J. 8, 3973 (1989)], was ligated into the Bam HI site of YEp351 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, Yeast 2, 163 (1986)].

24. J. Chant and I. Herskowitz, Cell 65, 1203 (1991).
25. B. W. Matthews, Acc. Chem. Res. 21, 333 (1988).
26. K. Kuchler, H. G. Dohlman, J. Thorner, J. Cell Biol. 120, 1203 (1993); R. Kolling and C. P. Hollenberg, EMBO J. 13, 3261 (1994); C. Berkower, D. Loayza, S. Michaelis, Mol. Biol. Cell 5, 1185 (1994).
27. A. Bender and J. R. Pringle, Proc. Natl. Acad. Sci. U.S.A 86, 9976 (1989); J. Chant, K. Corrado, J. R. Pringle, I. Herskowitz, Cell 65, 1213 (1991); I. S. Powers, E. Gonzales, T. Christensen, J. Cubert, D. Broek, ibid., p. 1225; H. O. Park, J. Chant, I. Herskowitz, Nature 365, 269 (1993); J. Chant, Trends Genet. 10, 328 (1994); _____ and J. R. Pringle, J. Cell Biol. 129, 751 (1995); J. Chant, M. Mischke, E. Mitchell, I. Herskowitz, J. R. Pringle, ibid., p. 767.
28. G. F. Sprague Jr., Methods. Enzymol. 194, 77 (1991).
29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr.
30. A W303 1A derivative, SY2625 (MATa ura3-1 leu2-3, 112 trp1-1 ade2-1 can1-100 sst1Δ mfa2Δ::FUS1-lacZ his3Δ::FUS1-HIS3), was the parent strain for the mutant search. SY2625 derivatives for the mating assays, secreted pheromone assays, and the pulse-chase experiments included the following strains: Y49 (sta22-1), Y115 (mfa1Δ::LEU2), Y142 (axl1::URA3), Y173 (axl1Δ::LEU2), Y220 (axl1::URA3 sta23Δ::URA3), Y221 (sta23Δ::URA3), Y231 (axl1Δ::LEU2 sta23Δ::LEU2), and Y233 (sta23Δ::LEU2). MATa derivatives of SY2625 included the following strains: Y199 (SY2625 made MATα), Y278 (sta22-1), Y195 (mfa1Δ::LEU2), Y196 (axl1Δ::LEU2), and Y197 (axl1::URA3). The EG123 (MATa leu2 ura3 trp1 can1 his4) genetic background was used to create a set of strains for analysis of bud site selection. EG123 derivatives included the following strains: Y175 (axl1Δ::LEU2), Y223 (axl1::URA3), Y234 (sta23Δ::LEU2), and Y272 (axl1Δ::LEU2 sta23Δ::LEU2). MATα derivatives of EG123 included the following strains: Y214 (EG123 made MATα) and Y293 (axl1Δ::LEU2). All strains were generated by means of standard genetic or molecular methods involving the appropriate constructs (23). In particular, the axl1 sta23 double mutant strains were created by crossing of the appropriate MATa sta23 and MATα axl1 mutants, followed by sporulation of the resultant diploid and isolation of the double mutant from nonparental di-type tetrads. Gene disruptions were confirmed with either PCR or Southern (DNA) analysis.
31. p129 is a YEp352 [J. E. Hill, A. M. Myers, T. J. Koerner, A. Tzagoloff, Yeast 2, 163 (1986)] plasmid containing a 5.5-kb Sal I fragment of pAXL1. p151 was derived from p129 by insertion of a linker at the Bgl II site within AXL1, which led to an in-frame insertion of the hemagglutinin (HA) epitope (DQYPYDVPDYA) (29) between amino acids 854 and 855 of the AXL1 prod-

uct. pC225 is a KS+ (Stratagene) plasmid containing a 0.5-kb Bam HI–Sst I fragment from pAXL1. Substitution mutations of the proposed active site of Axl1p were created with the use of pC225 and site-specific mutagenesis involving appropriate synthetic oligonucleotides [axl1-H68A, 5'-GTGCTCACAAAGCGCT-GCCAAACCGGC-3'; axl1-E71A, 5'-AAGAATCAT-GTGCGCACAAAGGTGCGC-3'; and axl1-E71D, 5'-AAGAATCATGTGATCACAAAGGTGCGC-3']. The mutations were confirmed by sequence analysis. After mutagenesis, the 0.4-kb Bam HI–Msc I fragment from the mutagenized pC225 plasmids was transferred into pAXL1 to create a set of pRS318 plasmids carrying different AXL1 alleles, p124 (axl1-H68A), p130 (axl1-E71A), and p132 (axl1-E71D). Similarly, a set of HA-tagged alleles carried on YEp352 were created after replacement of the p151 Bam HI–Msc I fragment, to generate p161 (axl1-E71A), p162 (axl1-

Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray

Mark Schena,* Dari Shalon,*† Ronald W. Davis, Patrick O. Brown‡

A high-capacity system was developed to monitor the expression of many genes in parallel. Microarrays prepared by high-speed robotic printing of complementary DNAs on glass were used for quantitative expression measurements of the corresponding genes. Because of the small format and high density of the arrays, hybridization volumes of 2 microliters could be used that enabled detection of rare transcripts in probe mixtures derived from 2 micrograms of total cellular messenger RNA. Differential expression measurements of 45 Arabidopsis genes were made by means of simultaneous, two-color fluorescence hybridization.

The temporal, developmental, topographical, histological, and physiological patterns in which a gene is expressed provide clues to its biological role. The large and expanding database of complementary DNA (cDNA) sequences from many organisms (1) presents the opportunity of defining these patterns at the level of the whole genome.

For these studies, we used the small flowering plant Arabidopsis thaliana as a model organism. Arabidopsis possesses many advantages for gene expression analysis, including the fact that it has the smallest genome of any higher eukaryote examined to date (2). Forty-five cloned Arabidopsis cDNAs (Table 1), including 14 complete sequences and 31 expressed sequence tags (ESTs), were used as gene-specific targets. We obtained the ESTs by selecting cDNA clones at random from an Arabidopsis cDNA library. Sequence analysis revealed that 28 of the 31 ESTs matched sequences

M. Schena and R. W. Davis, Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.
D. Shalon and P. O. Brown, Department of Biochemistry and Howard Hughes Medical Institute, Beckman Center, Stanford University Medical Center, Stanford, CA 94305, USA.

*These authors contributed equally to this work.
†Present address: Synteni, Palo Alto, CA 94303, USA.
‡To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

in the database (Table 1). Three additional cDNAs from other organisms served as controls in the experiments.

The 48 cDNAs, averaging ~1.0 kb, were amplified with the polymerase chain reaction (PCR) and deposited into individual wells of a 96-well microtiter plate. Each sample was duplicated in two adjacent wells to allow the reproducibility of the arraying and hybridization process to be tested. Samples from the microtiter plate were printed onto glass microscope slides in an area measuring 3.5 mm by 5.5 mm with the use of a high-speed arraying machine (3). The arrays were processed by chemical and heat treatment to attach the DNA sequences to the glass surface and denature them (3). Three arrays, printed in a single lot, were used for the experiments here. A single microtiter plate of PCR products provides sufficient material to print at least 500 arrays.

Fluorescent probes were prepared from total Arabidopsis mRNA (4) by a single round of reverse transcription (5). The Arabidopsis mRNA was supplemented with human acetylcholine receptor (AChR) mRNA at a dilution of 1:10,000 (w/w) before cDNA synthesis, to provide an internal standard for calibration (5). The resulting fluorescently labeled cDNA mixture was hybridized to an array at high stringency (6) and scanned

with a laser (3). A high-sensitivity scan gave signals that saturated the detector at nearly all of the *Arabidopsis* target sites (Fig. 1A). Calibration relative to the AChR mRNA standard (Fig. 1A) established a sensitivity limit of ~1:50,000. No detectable hybridization was observed to either the rat glucocorticoid receptor (Fig. 1A) or the yeast *TRP4* (Fig. 1A) targets even at the highest scanning sensitivity. A moderate-sensitivity scan

of the same array allowed linear detection of the more abundant transcripts (Fig. 1B). Quantitation of both scans revealed a range of expression levels spanning three orders of magnitude for the 45 genes tested (Table 2). RNA blots (7) for several genes (Fig. 2) corroborated the expression levels measured with the microarray to within a factor of 5 (Table 2).

Differential gene expression was investi-

gated with a simultaneous, two-color hybridization scheme, which served to minimize experimental variation inherent in the comparison of independent hybridizations. Fluorescent probes were prepared from two mRNA sources with the use of reverse transcriptase in the presence of fluorescein- and lissamine-labeled nucleotide analogs, respectively (5). The two probes were then mixed together in equal proportions, hybridized to a single array, and scanned separately for fluorescein and lissamine emission after independent excitation of the two fluorophores (3).

To test whether overexpression of a single gene could be detected in a pool of total *Arabidopsis* mRNA, we used a microarray to analyze a transgenic line overexpressing the single transcription factor *HAT4* (8). Fluorescent probes representing mRNA from wild-type and *HAT4*-transgenic plants were labeled with fluorescein and lissamine, respectively; the two probes were then mixed and hybridized to a single array. An intense hybridization signal was observed at the position of the *HAT4* cDNA in the lissamine-specific scan (Fig. 1D), but not in the fluorescein-specific scan of the same array (Fig. 1C). Calibration with AChR mRNA added to the fluorescein and lissamine cDNA synthesis reactions at dilutions of 1:10,000 (Fig. 1C) and 1:100 (Fig. 1D), respectively, revealed a 50-fold elevation of *HAT4* mRNA in the transgenic line relative to its abundance in wild-type plants (Table 2). This magnitude of *HAT4* overexpression matched that inferred from the Northern (RNA) analysis within a factor of 2 (Fig. 2 and Table 2). Expression of all the other genes monitored on the array differed by less than a factor of 5 between *HAT4*-transgenic and wild-type plants (Fig 1, C
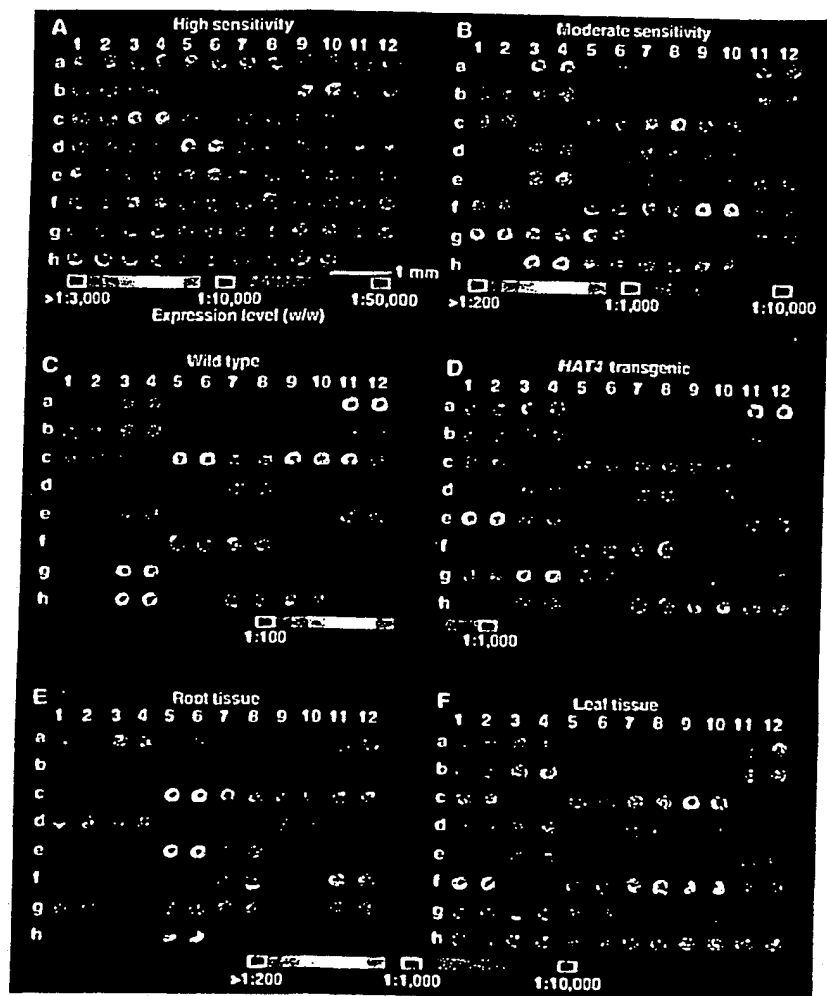

Fig. 1. Gene expression monitored with the use of cDNA microarrays. Fluorescent scans represented in pseudocolor correspond to hybridization intensities. Color bars were calibrated from the signal obtained with the use of known concentrations of human AChR mRNA in independent experiments. Numbers and letters on the axes mark the position of each cDNA. (A) High-sensitivity fluorescein scan after hybridization with fluorescein-labeled cDNA derived from wild-type plants. (B) Same array as in (A) but scanned at moderate sensitivity. (C and D) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from wild-type plants and lissamine-labeled cDNA from HAT4-transgenic plants. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNA from wild-type plants (C) and the lissamine fluorescence corresponding to mRNA from HAT4-transgenic plants (D). (E and F) A single array was probed with a 1:1 mixture of fluorescein-labeled cDNA from root tissue and lissamine-labeled cDNA from leaf tissue. The single array was then scanned successively to detect the fluorescein fluorescence corresponding to mRNAs expressed in roots (E) and the lissamine fluorescence corresponding to mRNAs expressed in leaves (F).
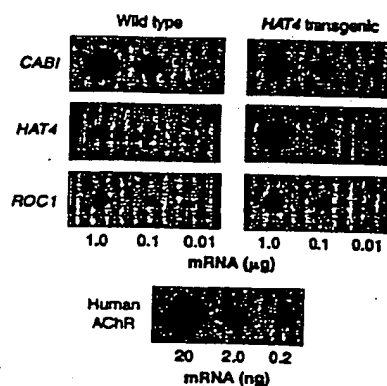

Fig. 2. Gene expression monitored with RNA (Northern) blot analysis. Designated amounts of mRNA from wild-type and HAT4-transgenic plants were spotted onto nylon membranes and probed with the cDNAs indicated. Purified human AChR mRNA was used for calibration.

and D, and Table 2). Hybridization of fluorescein-labeled glucocorticoid receptor cDNA (Fig. 1C) and lissamine-labeled TRP4 cDNA (Fig. 1D) verified the presence of the negative control targets and the lack of optical cross talk between the two fluorophores.

To explore a more complex alteration in expression patterns, we performed a second two-color hybridization experiment with fluorescein- and lissamine-labeled probes prepared from root and leaf mRNA, respectively. The scanning sensitivities for the two fluorophores were normalized by matching the signals resulting from AChR mRNA, which was added to both cDNA synthesis reactions at a dilution of 1:1000 (Fig. 1, E and F). A comparison of the scans revealed widespread differences in gene expression between root and leaf tissue (Fig. 1, E and F). The mRNA from the light-regulated CABl gene was ~500-fold more abundant in leaf (Fig. 1F) than in root tissue (Fig. 1E). The expression of 26 other genes differed between root and leaf tissue by more than a factor of 5 (Fig. 1, E and F).

The HAT4-transgenic line we examined has elongated hypocotyls, early flowering, poor germination, and altered pigmentation (8). Although changes in expression were observed for HAT4, large changes in expression were not observed for any of the other 44 genes we examined. This was somewhat surprising, particularly because comparative analysis of leaf and root tissue identified 27 differentially expressed genes. Analysis of an expanded set of genes may be required to identify genes whose expression changes upon HAT4 overexpression; alternatively, a comparison of mRNA populations from specific tissues of wild-type and HAT4-transgenic plants may allow identification of downstream genes.

At the current density of robotic printing, it is feasible to scale up the fabrication process to produce arrays containing 20,000 cDNA targets. At this density, a single array would be sufficient to provide gene-specific targets encompassing nearly the entire repertoire of expressed genes in the Arabidopsis genome (2). The availability of 20,274 ESTs from Arabidopsis (1, 9) would provide a rich source of templates for such studies.

The estimated 100,000 genes in the human genome (10) exceeds the number of Arabidopsis genes by a factor of 5 (2). This modest increase in complexity suggests that similar cDNA microarrays, prepared from the rapidly growing repertoire of human ESTs (1), could be used to determine the expression patterns of tens of thousands of human genes in diverse cell types. Coupling an amplification strategy to the reverse transcription reaction (11) could make it feasible to monitor expression even in minute tissue samples. A wide variety of acute and chronic physiological and pathological conditions might lead to characteristic changes in the patterns of gene expression in peripheral blood cells or other easily sampled tissues. In concert with cDNA microarrays for monitoring complex expression patterns, these tissues might therefore serve as sensitive in vivo sensors for clinical diagnosis. Microarrays of cDNAs could thus provide a useful link between human gene sequences and clinical medicine.

Table 1. Sequences contained on the cDNA microarray. Shown is the position, the known or putative function, and the accession number of each cDNA in the microarray (Fig. 1). All but three of the ESTs used in this study matched a sequence in the database. NADH, reduced form of nicotinamide adenine dinucleotide; ATPase, adenosine triphosphatase; GTP, guanosine triphosphate.

Position	cDNA	Function	Accession number
a1, 2	AChR	Human AChR	
a3, 4	EST3	Actin	H36236
a5, 6	EST6	NADH dehydrogenase	Z27010
a7, 8	AAC1	Actin 1	M20016
a9, 10	EST12	Unknown	U36594†
a11, 12	EST13	Actin	T45783
b1, 2	CABl	Chlorophyll a/b binding	M85150
b3, 4	EST17	Phosphoglycerate kinase	T44490
b5, 6	GA4	Gibberellic acid biosynthesis	L37126
b7, 8	EST19	Unknown	U36595†
b9, 10	GBF-1	G-box binding factor 1	X63894
b11, 12	EST23	Elongation factor	X52256
c1, 2	EST29	Aldolase	T04477
c3, 4	GBF-2	G-box binding factor 2	X63895
c5, 6	EST34	Chloroplast protease	R87034
c7, 8	EST35	Unknown	T14152
c9, 10	EST41	Catalase	T22720
c11, 12	rGR	Rat glucocorticoid receptor	M14053
d1, 2	EST42	Unknown	U36596†
d3, 4	EST45	ATPase	J04185
d5, 6	HAT1	Homeobox-leucine zipper 1	U09332
d7, 8	EST46	Light harvesting complex	T04063
d9, 10	EST49	Unknown	T76267
d11, 12	HAT2	Homeobox-leucine zipper 2	U09335
e1, 2	HAT4	Homeobox-leucine zipper 4	M90394
e3, 4	EST50	Phosphoribulokinase	T04344
e5, 6	HAT5	Homeobox-leucine zipper 5	M90416
e7, 8	EST51	Unknown	Z33675
e9, 10	HAT22	Homeobox-leucine zipper 22	U09336
e11, 12	EST52	Oxygen evolving	T21749
f1, 2	EST59	Unknown	Z34607
f3, 4	KNAT1	Knotted-like homeobox 1	U14174
f5, 6	EST60	RuBisCO small subunit	X14564
f7, 8	EST69	Translation elongation factor	T42799
f9, 10	PPH1	Protein phosphatase 1	U34803
f11, 12	EST70	Unknown	T44621
g1, 2	EST75	Chloroplast protease	T43698
g3, 4	EST78	Unknown	R65481
g5, 6	ROC1	Cyclophilin	L14844
g7, 8	EST82	GTP binding	X59152
g9, 10	EST83	Unknown	Z33795
g11, 12	EST84	Unknown	T45278
h1, 2	EST91	Unknown	T13832
h3, 4	EST96	Unknown	R64816
h5, 6	SAR1	Synaptobrevin	M90418
h7, 8	EST100	Light harvesting complex	Z18205
h9, 10	EST103	Light harvesting complex	X03909
h11, 12	TRP4	Yeast tryptophan biosynthesis	X04273

*Proprietary sequence of Stratagene (La Jolla, California). †No match in the database; novel EST.

Table 2. Gene expression monitoring by microarray and RNA blot analyses; tg, HAT4-transgenic. See Table 1 for additional gene information. Expression levels (w/w) were calibrated with the use of known amounts of human AChR mRNA. Values for the microarray were determined from microarray scans (Fig. 1); values for the RNA blot were determined from RNA blots (Fig. 2).

Gene	Expression level (w/w)	
	Microarray	RNA blot
CABl	1:48	1:83
CABl (tg)	1:120	1:150
HAT4	1:8300	1:6300
HAT4 (tg)	1:150	1:210
ROC1	1:1200	1:1800
ROC1 (tg)	1:260	1:1300

REFERENCES AND NOTES

1. The current EST database (dbEST release 091495) from the National Center for Biotechnology Information (Bethesda, MD) contains a total of 322,225 entries, including 255,645 from the human genome and 21,044 from *Arabidopsis*. Access is available via the World Wide Web (http://www.ncbi.nlm.nih.gov).
2. E. M. Meyerowitz and R. E. Pruitt, *Science* 229, 1214 (1985); R. E. Pruitt and E. M. Meyerowitz, *J. Mol. Biol.* 187, 169 (1986); L. Hwang *et al.*, *Plant J.* 1, 367 (1991); P. Jarvis *et al.*, *Plant Mol. Biol.* 24, 685 (1994); L. Le Guen *et al.*, *Mol. Gen. Genet.* 245, 390 (1994).
3. D. Shalon, thesis, Stanford University (1995); _____ and P. O. Brown, in preparation. Microarrays were fabricated on poly-L-lysine–coated microscope slides (Sigma) with a custom-built arraying machine fitted with one printing tip. The tip loaded 1 μl of PCR product (0.5 mg/ml) from 96-well microtiter plates and deposited ~0.005 μl per slide on 40 slides at a spacing of 500 μm. The printed slides were rehydrated for 2 hours in a humid chamber, snap-dried at 100°C for 1 min, rinsed in 0.1% SDS, and treated with 0.05% succinic anhydride prepared in buffer consisting of 50% 1-methyl-2-pyrrolidinone and 50% boric acid. The cDNA on the slides was denatured in distilled water for 2 min at 90°C immediately before use. Microarrays were scanned with a laser fluorescent scanner that contained a computer-controlled XY stage and a microscope objective. A mixed gas, multiline laser allowed sequential excitation of the two fluorophores. Emitted light was spilt according to wavelength and detected with two photomultiplier tubes. Signals were read into a PC with the use of a 12-bit analog-to-digital board. Additional details of microarray fabrication and use may be obtained by means of e-mail (pbrown@cmgm. stanford.edu).
4. F. M. Ausubel *et al.*, Eds., *Current Protocols in Molecular Biology* (Greene & Wiley Interscience, New York, 1994), pp. 4.3.1–4.3.4.
5. Polyadenylated [poly(A)⁺] mRNA was prepared from total RNA with the use of Oligotex-dT resin (Qiagen). Reverse transcription (RT) reactions were carried out with a StrataScript RT-PCR kit (Stratagene) modified as follows: 50-μl reactions contained 0.1 μg/μl of *Arabidopsis* mRNA, 0.1 ng/μl of human AChR mRNA, 0.05 μg/μl of oligo(dT) (21-mer), 1× first strand buffer, 0.03 U/μl of ribonuclease block, 500 μM deoxyadenosine triphosphate (dATP), 500 μM deoxyguanosine triphosphate, 500 μM dTTP, 40 μM deoxycytosine triphosphate (dCTP), 40 μM fluorescein-12-dCTP (or lissamine-5-dCTP), and 0.03 U/μl of StrataScript reverse transcriptase. Reactions were incubated for 60 min at 37°C, precipitated with ethanol, and resuspended in 10 μl of TE (10 mM tris-HCl and 1 mM EDTA, pH 8.0). Samples were then heated for 3 min at 94°C and chilled on ice. The RNA was degraded by adding 0.25 μl of 10 N NaOH followed by a 10-min incubation at 37°C. The samples were neutralized by addition of 2.5 μl of 1 M tris-Cl (pH 8.0) and 0.25 μl of 10 N HCl and precipitated with ethanol. Pellets were washed with 70% ethanol, dried to completion in a speedvac, resuspended in 10 μl of H₂O, and reduced to 3.0 μl in a speedvac. Fluorescent nucleotide analogs were obtained from New England Nuclear (DuPont).
6. Hybridization reactions contained 1.0 μl of fluorescent cDNA synthesis product (5) and 1.0 μl of hybridization buffer [10× saline sodium citrate (SSC) and 0.2% SDS]. The 2.0-μl probe mixtures were aliquoted onto the microarray surface and covered with cover slips (12 mm round). Arrays were transferred to a hybridization chamber (3) and incubated for 18 hours at 65°C. Arrays were washed for 5 min at room temperature (25°C) in low-stringency wash buffer (1× SSC and 0.1% SDS), then for 10 min at room temperature in high-stringency wash buffer (0.1× SSC and 0.1% SDS). Arrays were scanned in 0.1× SSC with the use of a fluorescence laser-scanning device (3).
7. Samples of poly(A)⁺ mRNA (4, 5) were spotted onto nylon membranes (Nytran) and crosslinked with ultraviolet light with the use of a Stratalinker 1800 (Stratagene). Probes were prepared by random priming with the use of a Prime-It II kit (Stratagene) in the presence of [³²P]dATP. Hybridizations were carried out according to the instructions of the manufacturer. Quantitation was performed on a Phosphorimager (Molecular Dynamics).
8. M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 89, 3894 (1992); M. Schena, A. M. Lloyd, R. W. Davis, *Genes Dev.* 7, 367 (1993); M. Schena and R. W. Davis, *Proc. Natl. Acad. Sci. U.S.A.* 91, 8393 (1994).
9. H. Hofte *et al.*, *Plant J.* 4, 1051 (1993); T. Newman *et al.*, *Plant Physiol.* 106, 1241 (1994).
10. N. E. Morton, *Proc. Natl. Acad. Sci. U.S.A.* 88, 7474 (1991); E. D. Green and R. H. Waterston, *J. Am. Med. Assoc.* 266, 1966 (1991); C. Bellanne-Chantelot, *Cell* 70, 1059 (1992); D. R. Cox *et al.*, *Science* 265, 2031 (1994).
11. E. S. Kawasaki *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* 85, 5698 (1988).
12. The laser fluorescent scanner was designed and fabricated in collaboration with S. Smith of Stanford University. Scanner and analysis software were developed by R. X. Xia. The succinic anhydride reaction was suggested by J. Mulligan and J. Van Ness of Darwin Molecular Corporation. Thanks to S. Theologis, C. Somerville, K. Yamamoto, and members of the laboratories of R.W.D. and P.O.B. for critical comments. Supported by the Howard Hughes Medical Institute and by grants from NIH [R21HG00450] (P.O.B.) and R37AG00198 (R.W.D.)] and from NSF (MCB9106011) (R.W.D.) and by an NSF graduate fellowship (D.S.). P.O.B. is an assistant investigator of the Howard Hughes Medical Institute.

11 August 1995; accepted 22 September 1995

Gene Therapy in Peripheral Blood Lymphocytes and Bone Marrow for ADA⁻ Immunodeficient Patients

Claudio Bordignon,* Luigi D. Notarangelo, Nadia Nobili, Giuliana Ferrari, Giulia Casorati, Paola Panina, Evelina Mazzolari, Daniela Maggioni, Claudia Rossi, Paolo Servida, Alberto G. Ugazio, Fulvio Mavilio

Adenosine deaminase (ADA) deficiency results in severe combined immunodeficiency, the first genetic disorder treated by gene therapy. Two different retroviral vectors were used to transfer ex vivo the human ADA minigene into bone marrow cells and peripheral blood lymphocytes from two patients undergoing exogenous enzyme replacement therapy. After 2 years of treatment, long-term survival of T and B lymphocytes, marrow cells, and granulocytes expressing the transferred ADA gene was demonstrated and resulted in normalization of the immune repertoire and restoration of cellular and humoral immunity. After discontinuation of treatment, T lymphocytes, derived from transduced peripheral blood lymphocytes, were progressively replaced by marrow-derived T cells in both patients. These results indicate successful gene transfer into long-lasting progenitor cells, producing a functional multilineage progeny.

Severe combined immunodeficiency associated with inherited deficiency of ADA (1) is usually fatal unless affected children are kept in protective isolation or the immune system is reconstituted by bone marrow transplantation from a human leukocyte antigen (HLA)–identical sibling donor (2). This is the therapy of choice, although it is available only for a minority of patients. In recent years, other forms of therapy have been developed, including transplants from haploidentical donors (3, 4), exogenous enzyme replacement (5), and somatic-cell gene therapy (6–9).

We previously reported a preclinical model in which ADA gene transfer and expression

successfully restored immune functions in human ADA-deficient (ADA⁻) peripheral blood lymphocytes (PBLs) in immunodeficient mice in vivo (10, 11). On the basis of these preclinical results, the clinical application of gene therapy for the treatment of ADA⁻ SCID (severe combined immunodeficiency disease) patients who previously failed exogenous enzyme replacement therapy was approved by our Institutional Ethical Committees and by the Italian National Committee for Bioethics (12). In addition to evaluating the safety and efficacy of the gene therapy procedure, the aim of the study was to define the relative role of PBLs and hematopoietic stem cells in the long-term reconstitution of immune functions after retroviral vector–mediated ADA gene transfer. For this purpose, two structurally identical vectors expressing the human ADA complementary DNA (cDNA), distinguishable by the presence of alternative restriction sites in a nonfunctional region of the viral long-terminal repeat (LTR), were used to transduce PBLs and bone marrow (BM) cells independently. This procedure allowed identification of the origin of

C. Bordignon, N. Nobili, G. Ferrari, D. Maggioni, C. Rossi, P. Servida, F. Mavilio, Telethon Gene Therapy Program for Genetic Diseases, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
L. D. Notarangelo, E. Mazzolari, A. G. Ugazio, Department of Pediatrics, University of Brescia Medical School, Brescia, Italy.
G. Casorati, Unità di Immunochimica, DIBIT, Istituto Scientifico H. S. Raffaele, Milan, Italy.
P. Panina, Roche Milano Ricerche, Milan, Italy.

*To whom correspondence should be addressed.

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(54) Title: METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

(57) Abstract

A method and apparatus for forming microarrays of biological samples on a support are disclosed. The method involves dispensing a known volume of a reagent at each of a selected array position, by tapping a capillary dispenser on the support under conditions effective to draw a defined volume of liquid onto the support. The apparatus is designed to produce a microarray of such regions in an automated fashion.

METHOD AND APPARATUS FOR FABRICATING MICROARRAYS OF BIOLOGICAL SAMPLES

Field of the Invention

5   This invention relates to a method and apparatus for fabricating microarrays of biological samples for large scale screening assays, such as arrays of DNA samples to be used in DNA hybridization assays for genetic research and diagnostic applications.

10

References

Abouzied, et al., *Journal of AOAC International* 77(2):495-500 (1994).

Bohlander, et al., *Genomics* 13:1322-1324 (1992).

15   Drmanac, et al., *Science* 260:1649-1652 (1993).

Fodor, et al., *Science* 251:767-773 (1991).

Khrapko, et al., *DNA Sequence* 1:375-388 (1991).

Kuriyama, et al., AN ISFET BIOSENSOR, APPLIED BIOSENSORS (Donald Wise, Ed.), Butterworths, pp. 93-114 (1989).

20   Lehrach, et al., HYBRIDIZATION FINGERPRINTING IN GENOME MAPPING AND SEQUENCING, GENOME ANALYSIS, VOL 1 (Davies and Tilgham, Eds.), Cold Spring Harbor Press, pp. 39-81 (1990).

Maniatis, et al., MOLECULAR CLONING, A LABORATORY

25 MANUAL, Cold Spring Harbor Press (1989).

Nelson, t al., *Nature Genetics* 4:11-18 (1993).

2

Pirrung, et al., U.S. Patent No. 5,143,854 (1992).

Riles, et al., Genetics 134:81-150 (1993).

Schena, M. et al., Proc. Nat. Acad. Sci. USA 89:3894-3898 (1992).

5   Southern, et al., Genomics 13:1008-1017 (1992).

Background of the Invention

A variety of methods are currently available for making arrays of biological macromolecules, such as
10   arrays of nucleic acid molecules or proteins. One method for making ordered arrays of DNA on a porous membrane is a "dot blot" approach. In this method, a vacuum manifold transfers a plurality, e.g., 96, aqueous samples of DNA from 3 millimeter diameter wells
15   to a porous membrane. A common variant of this procedure is a "slot-blot" method in which the wells have highly-elongated oval shapes.

The DNA is immobilized on the porous membrane by baking the membrane or exposing it to UV radiation.
20   This is a manual procedure practical for making one array at a time and usually limited to 96 samples per array. "Dot-blot" procedures are therefore inadequate for applications in which many thousand samples must be determined.

25   A more efficient technique employed for making ordered arrays of genomic fragments uses an array of pins dipped into the wells, e.g., the 96 wells of a microtitre plate, for transferring an array of samples to a substrate, such as a porous membrane. One array
30   includes pins that are designed to spot a membrane in a staggered fashion, for creating an array of 9216 spots in a 22 × 22 cm area (Lehrach, et al., 1990). A limitation with this approach is that the volume of DNA spotted in each pixel of each array is highly variable.

3

In addition, the number of arrays that can be made with each dipping is usually quite small.

An alternate method of creating ordered arrays of nucleic acid sequences is described by Pirrung, et al.
5   (1992), and also by Fodor, et al. (1991). The method involves synthesizing different nucleic acid sequences at different discrete regions of a support. This method employs elaborate synthetic schemes, and is generally limited to relatively short nucleic acid
10   sample, e.g., less than 20 bases. A related method has been described by Southern, et al. (1992).

Khrapko, et al. (1991) describes a method of making an oligonucleotide matrix by spotting DNA onto a thin layer of polyacrylamide. The spotting is done
15   manually with a micropipette.

None of the methods or devices described in the prior art are designed for mass fabrication of microarrays characterized by (i) a large number of micro-sized assay regions separated by a distance of
20   50-200 microns or less, and (ii) a well-defined amount, typically in the picomole range, of analyte associated with each region of the array.

Furthermore, current technology is directed at performing such assays one at a time to a single array
25   of DNA molecules. For example, the most common method for performing DNA hybridizations to arrays spotted onto porous membrane involves sealing the membrane in a plastic bag (Maniatas, et al., 1989) or a rotating glass cylinder (Robbins Scientific) with the labeled
30   hybridization probe inside the sealed chamber. For arrays made on non-porous surfaces, such as a microscope slide, each array is incubated with the labeled hybridization probe sealed under a coverslip. These techniques require a separate sealed chamber for

4

each array which makes the screening and handling of
many such arrays inconvenient and time intensive.

Abouzied, et al. (1994) describes a method of
printing horizontal lines of antibodies on a
5   nitrocellulose membrane and separating regions of the
membrane with vertical stripes of a hydrophobic
material. Each vertical stripe is then reacted with a
different antigen and the reaction between the
immobilized antibody and an antigen is detected using a
10   standard ELISA colorimetric technique. Abouzied's
technique makes it possible to screen many one-
dimensional arrays simultaneously on a single sheet of
nitrocellulose. Abouzied makes the nitrocellulose
somewhat hydrophobic using a line drawn with PAP Pen
15   (Research Products International). However Abouzied
does not describe a technology that is capable of
completely sealing the pores of the nitrocellulose. The
pores of the nitrocellulose are still physically open
and so the assay reagents can leak through the
20   hydrophobic barrier during extended high temperature
incubations or in the presence of detergents which
makes the Abouzied technique unacceptable for DNA
hybridization assays.

Porous membranes with printed patterns of
25   hydrophilic/hydrophobic regions exist for applications
such as ordered arrays of bacteria colonies. QA Life
Sciences (San Diego CA) makes such a membrane with a
grid pattern printed on it. However, this membrane has
the same disadvantage as the Abouzied technique since
30   reagents can still flow between the gridded arrays
making them unusable for separate DNA hybridization
assays.

Pall Corporation make a 96-well plate with a
porous filter heat sealed to the bottom of the plate.
35   These plates are capable of containing different

5

reagents in each well without cross-contamination.
However, each well is intended to hold only one target
element whereas the invention described here makes a
microarray of many biomolecules in each subdivided
5   region of the solid support. Furthermore, the 96 well
plates are at least 1 cm thick and prevent the use of
the device for many colorimetric, fluorescent and
radioactive detection formats which require that the
membrane lie flat against the detection surface. The
10   invention described here requires no further processing
after the assay step since the barriers elements are
shallow and do not interfere with the detection step
thereby greatly increasing convenience.

Hyseq Corporation has described a method of making
15   an "array of arrays" on a non-porous solid support for
use with their sequencing by hybridization technique.
The method described by Hyseq involves modifying the
chemistry of the solid support material to form a
hydrophobic grid pattern where each subdivided region
20   contains a microarray of biomolecules. Hyseq's flat
hydrophobic pattern does not make use of physical
blocking as an additional means of preventing cross
contamination.

25   **Summary of the Invention**

The invention includes, in one aspect, a method of
forming a microarray of analyte-assay regions on a
solid support, where each region in the array has a
known amount of a selected, analyte-specific reagent.
30   The method involves first loading a solution of a
selected analyte-specific reagent in a reagent-
dispensing device having an elongate capillary channel
(i) formed by spaced-apart, coextensive elongate
members, (ii) adapted to hold a quantity of the reagent
35   solution and (iii) having a tip r gion at which aqueous

6

solution in the channel forms a meniscus. The channel
is preferably formed by a pair of spaced-apart tapered
elements.

The tip of the dispensing device is tapped against
5   a solid support at a defined position on the support
surface with an impulse effective to break the meniscus
in the capillary channel deposit a selected volume of
solution on the surface, preferably a selected volume
in the range 0.01 to 100 nl. The two steps are
10   repeated until the desired array is formed.

The method may be practiced in forming a plurality
of such arrays, where the solution-depositing step is
are applied to a selected position on each of a
plurality of solid supports at each repeat cycle.

15   The dispensing device may be loaded with a new
solution, by the steps of (i) dipping the capillary
channel of the device in a wash solution, (ii) removing
wash solution drawn into the capillary channel, and
(iii) dipping the capillary channel into the new
20   reagent solution.

Also included in the invention is an automated
apparatus for forming a microarray of analyte-assay
regions on a plurality of solid supports, where each
region in the array has a known amount of a selected,
25   analyte-specific reagent. The apparatus has a holder
for holding, at known positions, a plurality of planar
supports, and a reagent dispensing device of the type
described above.

The apparatus further includes positioning
30   structure for positioning the dispensing device at a
selected array position with respect to a support in
said holder, and dispensing structure for moving the
dispensing device into tapping engagement against a
support with a selected impulse effective to deposit a

7

selected volume on the support, e.g., a selected volume
in the volume range 0.01 to 100 nl.

The positioning and dispensing structures are
controlled by a control unit in the apparatus. The
5   unit operates to (i) place the dispensing device at a
loading station, (ii) move the capillary channel in the
device into a selected reagent at the loading station,
to load the dispensing device with the reagent, and
(iii) dispense the reagent at a defined array position
10   on each of the supports on said holder. The unit may
further operate, at the end of a dispensing cycle, to
wash the dispensing device by (i) placing the
dispensing device at a washing station, (ii) moving the
capillary channel in the device into a wash fluid, to
15   load the dispensing device with the fluid, and (iii)
remove the wash fluid prior to loading the dispensing
device with a fresh selected reagent.

The dispensing device in the apparatus may be one
of a plurality of such devices which are carried on the
20   arm for dispensing different analyte assay reagents at
selected spaced array positions.

In another aspect, the invention includes a
substrate with a surface having a microarray of at
least 10^3 distinct polynucleotide or polypeptide
25   biopolymers in a surface area of less than about 1 cm^2.
Each distinct biopolymer (i) is disposed at a separate,
defined position in said array, (ii) has a length of at
least 50 subunits, and (iii) is present in a defined
amount between about 0.1 femtomoles and 100 nanomoles.

30   In one embodiment, the surface is glass slide
surface coated with a polycationic polymer, such as
polylysine, and the biopolymers are polynucleotides.
In another embodiment, the substrate has a water-
impermeable backing, a water-permeable film formed on

8

the backing, and a grid formed on the film. The grid
is composed of intersecting water-impervious grid
elements extending from said backing to positions
raised above the surface of said film, and partitions
5   the film into a plurality of water-impervious cells. A
biopolymer array is formed within each well.

    More generally, there is provided a substrate for
use in detecting binding of labeled polynucleotides to
one or more of a plurality different-sequence,
10   immobilized polynucleotides. The substrate includes,
in one aspect, a glass support, a coating of a
polycationic polymer, such as polylysine, on said
surface of the support, and an array of distinct
polynucleotides electrostatically bound non-covalently
15   to said coating, where each distinct biopolymer is
disposed at a separate, defined position in a surface
array of polynucleotides.

    In another aspect, the substrate includes a water-
impermeable backing, a water-permeable film formed on
20   the backing, and a grid formed on the film, where the
grid is composed of intersecting water-impervious grid
elements extending from the backing to positions raised
above the surface of the film, forming a plurality of
cells. A biopolymer array is formed within each cell.

25   Also forming part of the invention is a method of
detecting differential expression of each of a
plurality of genes in a first cell type, with respect
to expression of the same genes in a second cell type.
In practicing the method, there is first produced
30   fluorescent-labeled cDNA's from mRNA's isolated from
the two cells types, where the cDNA'S from the first
and second cells are labeled with first and second
different fluorescent reporters.

    A mixture of the labeled cDNA's from the two cell
35   types is added to an array of polynucleotides

9

representing a plurality of known genes derived from
the two cell types, under conditions that result in
hybridization of the cDNA's to complementary-sequence
polynucleotides in the array. The array is then
5   examined by fluorescence under fluorescence excitation
conditions in which (i) polynucleotides in the array
that are hybridized predominantly to cDNA's derived
from one of the first and second cell types give a
distinct first or second fluorescence emission color,
10 respectively, and (ii) polynucleotides in the array
that are hybridized to substantially equal numbers of
cDNA's derived from the first and second cell types
give a distinct combined fluorescence emission color,
respectively. The relative expression of known genes
15 in the two cell types can then be determined by the
observed fluorescence emission color of each spot.
        These and other objects and features of the
invention will become more fully apparent when the
following detailed description of the invention is read
20 in conjunction with the accompanying figures.

                Brief Description of the Drawings
        Fig. 1 is a side view of a reagent-dispensing
device having a open-capillary dispensing head
25 constructed for use in one embodiment of the invention;
        Figs. 2A-2C illustrate steps in the delivery of a
fixed-volume bead on a hydrophobic surface employing
the dispensing head from Fig. 1, in accordance with one
embodiment of the method of the invention;
30   Fig. 3 shows a portion of a two-dimensional array
of analyte-assay regions constructed according to the
method of the invention;
        Fig. 4 is a planar view showing components of an
automated apparatus for forming arrays in accordance
35 with the invention.

10

Fig. 5 shows a fluorescent image of an actual 20 ×
20 array of 400 fluorescently-labeled DNA samples
immobilized on a poly-l-lysine coated slide, where the
total area covered by the 400 element array is 16
5   square millimeters;

Fig. 6 is a fluorescent image of a 1.8 cm × 1.8 cm
microarray containing lambda clones with yeast inserts,
the fluorescent signal arising from the hybridization
to the array with approximately half the yeast genome
10 labeled with a green fluorophore and the other half
with a red fluorophore;

Fig. 7 shows the translation of the hybridization
image of Fig. 6 into a karyotype of the yeast genome,
where the elements of Fig.-6 microarray contain yeast
15 DNA sequences that have been previously physically
mapped in the yeast genome;

Fig. 8 show a fluorescent image of a 0.5 cm × 0.5
cm microarray of 24 cDNA clones, where the microarray
was hybridized simultaneously with total cDNA from wild
20 type Arabidopsis plant labeled with a green fluorophore
and total cDNA from a transgenic Arabidopsis plant
labeled with a red fluorophore, and the arrow points to
the cDNA clone representing the gene introduced into
the transgenic Arabidopsis plant;

25 Fig. 9 shows a plan view of substrate having an
array of cells formed by barrier elements in the form
of a grid;

Fig. 10 shows an enlarged plan view of one of the
cells in the substrate in Fig. 9, showing an array of
30 polynucleotide regions in the cell;

Fig. 11 is an enlarged sectional view of the
substrate in Fig. 9, taken along a section line in that
figure; and

Fig. 12 is a scanned image of a 3 cm × 3 cm
35 nitrocellulose solid support containing four identical

11

arrays of M13 clones in each of four quadrants, where
each quadrant was hybridized simultaneously to a
different oligonucleotide using an open face
hybridization method.

5

Detailed Description of the Invention

I.   Definitions

Unless indicated otherwise, the terms defined
below have the following meanings:

10   "Ligand" refers to one member of a ligand/anti-
ligand binding pair. The ligand may be, for example,
one of the nucleic acid strands in a complementary,
hybridized nucleic acid duplex binding pair; an
effector molecule in an effector/receptor binding pair;

15   or an antigen in an antigen/antibody or
antigen/antibody fragment binding pair.

"Antiligand" refers to the opposite member of a
ligand/anti-ligand binding pair. The antiligand may be
the other of the nucleic acid strands in a

20   complementary, hybridized nucleic acid duplex binding
pair; the receptor molecule in an effector/receptor
binding pair; or an antibody or antibody fragment
molecule in antigen/antibody or antigen/antibody
fragment binding pair, respectively.

25   "Analyte" or "analyte molecule" refers to a
molecule, typically a macromolecule, such as a
polynucleotide or polypeptide, whose presence, amount,
and/or identity are to be determined. The analyte is
one member of a ligand/anti-ligand pair.

30   "Analyte-specific assay reagent" refers to a
molecule effective to bind specifically to an analyte
molecule. The reagent is the opposite member of a
ligand/anti-ligand binding pair.

An "array of regions on a solid support" is a

35   linear or two-dimensional array of preferably discrete

12

regions, each having a finite area, formed on the
surface of a solid support.

A "microarray" is an array of regions having a
density of discrete regions of at least about 100/cm^2,
5   and preferably at least about 1000/cm^2. The regions in
a microarray have typical dimensions, e.g., diameters,
in the range of between about 10-250 μm, and are
separated from other regions in the array by about the
same distance.

10   A support surface is "hydrophobic" if a aqueous-
medium droplet applied to the surface does not spread
out substantially beyond the area size of the applied
droplet. That is, the surface acts to prevent
spreading of the droplet applied to the surface by
15   hydrophobic interaction with the droplet.

A "meniscus" means a concave or convex surface
that forms on the bottom of a liquid in a channel as a
result of the surface tension of the liquid.

"Distinct biopolymers", as applied to the
20   biopolymers forming a microarray, means an array member
which is distinct from other array members on the basis
of a different biopolymer sequence, and/or different
concentrations of the same or distinct biopolymers,
and/or different mixtures of distinct or different-
25   concentration biopolymers. Thus an array of "distinct
polynucleotides" means an array containing, as its
members, (i) distinct polynucleotides, which may have a
defined amount in each member, (ii) different, graded
concentrations of given-sequence polynucleotides,
30   and/or (iii) different-composition mixtures of two or
more distinct polynucleotides.

"Cell type" means a cell from a given source,
e.g., a tissue, or organ, or a cell in a given state of

13

differentiation, or a cell associated with a given pathology or genetic makeup.

II. Method of Microarray Formation

5   This section describes a method of forming a microarray of analyte-assay regions on a solid support or substrate, where each region in the array has a known amount of a selected, analyte-specific reagent.

        Fig. 1 illustrates, in a partially schematic view,
10   a reagent-dispensing device 10 useful in practicing the method. The device generally includes a reagent dispenser 12 having an elongate open capillary channel 14 adapted to hold a quantity of the reagent solution, such as indicated at 16, as will be described below.
15   The capillary channel is formed by a pair of spaced-apart, coextensive, elongate members 12a, 12b which are tapered toward one another and converge at a tip or tip region 18 at the lower end of the channel. More generally, the open channel is formed by at least two
20   elongate, spaced-apart members adapted to hold a quantity of reagent solutions and having a tip region at which aqueous solution in the channel forms a meniscus, such as the concave meniscus illustrated at 20 in Fig. 2A. The advantages of the open channel
25   construction of the dispenser are discussed below.

        With continued reference to Fig. 1, the dispenser device also includes structure for moving the dispenser rapidly toward and away from a support surface, for effecting deposition of a known amount of solution in
30   the dispenser on a support, as will be described below with reference to Figs. 2A-2C. In the embodiment shown, this structure includes a solenoid 22 which is activatable to draw a solenoid piston 24 rapidly downwardly, then release the piston, e.g., under spring
35   bias, to a normal, raised position, as shown. The

14

dispenser is carried on the piston by a connecting
member 26, as shown. The just-described moving
structure is also referred to herein as dispensing
means for moving the dispenser into engagement with a
5   solid support, for dispensing a known volume of fluid
on the support.

The dispensing device just described is carried on
an arm 28 that may be moved either linearly or in an x-
y plane to position the dispenser at a selected
10   deposition position, as will be described.

Figs. 2A-2C illustrate the method of depositing a
known amount of reagent solution in the just-described
dispenser on the surface of a solid support, such as
the support indicated at 30. The support is a polymer,
15   glass, or other solid-material support having a surface
indicated at 31.

In one general embodiment, the surface is a
relatively hydrophilic, *i.e.*, wettable surface, such as
a surface having native, bound or covalently attached
20   charged groups. On such surface described below is a
glass surface having an absorbed layer of a
polycationic polymer, such as poly-l-lysine.

In another embodiment, the surface has or is
formed to have a relatively hydrophobic character,
25   *i.e.*, one that causes aqueous medium deposited on the
surface to bead. A variety of known hydrophobic
polymers, such as polystyrene, polypropylene, or
polyethylene have desired hydrophobic properties, as do
glass and a variety of lubricant or other hydrophobic
30   films that may be applied to the support surface.

Initially, the dispenser is loaded with a selected
analyte-specific reagent solution, such as by dipping
the dispenser tip, after washing, into a solution of
the reagent, and allowing filling by capillary flow
35   into the dispenser channel. The dispenser is now moved

15

to a selected position with respect to a support
surface, placing the dispenser tip directly above the
support-surface position at which the reagent is to be
deposited. This movement takes place with the
5   dispenser tip in its raised position, as seen in Fig.
2A, where the tip is typically at least several 1-5 mm
above the surface of the substrate.

With the dispenser so positioned, solenoid 22 is
now activated to cause the dispenser tip to move
10   rapidly toward and away from the substrate surface,
making momentary contact with the surface, in effect,
tapping the tip of the dispenser against the support
surface. The tapping movement of the tip against the
surface acts to break the liquid meniscus in the tip
15   channel, bringing the liquid in the tip into contact
with the support surface. This, in turn, produces a
flowing of the liquid into the capillary space between
the tip and the surface, acting to draw liquid out of
the dispenser channel, as seen in Fig. 2B.
20   Fig. 2C shows flow of fluid from the tip onto the
support surface, which in this case is a hydrophobic
surface. The figure illustrates that liquid continues
to flow from the dispenser onto the support surface
until it forms a liquid bead 32. At a given bead size,
25   i.e., volume, the tendency of liquid to flow onto the
surface will be balanced by the hydrophobic surface
interaction of the bead with the support surface, which
acts to limit the total bead area on the surface, and
by the surface tension of the droplet, which tends
30   toward a given bead curvature. At this point, a given
bead volume will have formed, and continued contact of
the dispenser tip with the bead, as the dispenser tip
is being withdrawn, will have little or no effect on
bead volume.

16

        For liquid-dispensing on a more hydrophilic
surface, the liquid will have less of a tendency to
bead, and the dispensed volume will be more sensitive
to the total dwell time of the dispenser tip in the
5   immediate vicinity of the support surface, e.g., the
positions illustrated in Figs. 2B and 2C.
        The desired deposition volume, i.e., bead volume,
formed by this method is preferably in the range 2 pl
(picoliters) to 2 nl (nanoliters), although volumes as
10   high as 100 nl or more may be dispensed. It will be
appreciated that the selected dispensed volume will
depend on (i) the "footprint" of the dispenser tip,
i.e., the size of the area spanned by the tip, (ii) the
hydrophobicity of the support surface, and (iii) the
15   time of contact with and rate of withdrawal of the tip
from the support surface. In addition, bead size may
be reduced by increasing the viscosity of the medium,
effectively reducing the flow time of liquid from the
dispenser onto the support surface. The drop size may
20   be further constrained by depositing the drop in a
hydrophilic region surrounded by a hydrophobic grid
pattern on the support surface.
        In a typical embodiment, the dispenser tip is
tapped rapidly against the support surface, with a
25   total residence time in contact with the support of
less than about 1 msec, and a rate of upward travel
from the surface of about 10 cm/sec.
        Assuming that the bead that forms on contact with
the surface is a hemispherical bead, with a diameter
30   approximately equal to the width of the dispenser tip,
as shown in Fig. 2C, the volume of the bead formed in
relation to dispenser tip width (d) is given in Table 1
below. As seen, the volume of the bead ranges between
2 pl to 2 nl as the width size is increased from about
35   20 to 200 μm.

17

Table 1

d	Volume (nl)
20 μm	2×10^{-3}
50 μm	3.1×10^{-2}
100 μm	2.5×10^{-1}
200 μm	2

At a given tip size, bead volume can be reduced in a controlled fashion by increasing surface hydrophobicity, reducing time of contact of the tip with the surface, increasing rate of movement of the tip away from the surface, and/or increasing the viscosity of the medium. Once these parameters are fixed, a selected deposition volume in the desired pl to nl range can be achieved in a repeatable fashion.

After depositing a bead at one selected location on a support, the tip is typically moved to a corresponding position on a second support, a droplet is deposited at that position, and this process is repeated until a liquid droplet of the reagent has been deposited at a selected position on each of a plurality of supports.

The tip is then washed to remove the reagent liquid, filled with another reagent liquid and this reagent is now deposited at each another array position on each of the supports. In one embodiment, the tip is washed and refilled by the steps of (i) dipping the capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

From the foregoing, it will be appreciated that the tweezers-like, open-capillary dispenser tip

18

provides the advantages that (i) the open channel of
the tip facilitates rapid, efficient washing and drying
before reloading the tip with a new reagent, (ii)
passive capillary action can load the sample directly
5   from a standard microwell plate while retaining
sufficient sample in the open capillary reservoir for
the printing of numerous arrays, (iii) open capillaries
are less prone to clogging than closed capillaries, and
(iv) open capillaries do not require a perfectly faced
10   bottom surface for fluid delivery.

A portion of a microarray 36 formed on the surface
38 of a solid support 40 in accordance with the method
just described is shown in Fig. 3. The array is formed
of a plurality of analyte-specific reagent regions,
15   such as regions 42, where each region may include a
different analyte-specific reagent. As indicated
above, the diameter of each region is preferably
between about 20-200 μm. The spacing between each
region and its closest (non-diagonal) neighbor,
20   measured from center-to-center (indicated at 44), is
preferably in the range of about 20-400 μm. Thus, for
example, an array having a center-to-center spacing of
about 250 μm contains about 40 regions/cm or 1,600
regions/cm^2. After formation of the array, the support
25   is treated to evaporate the liquid of the droplet
forming each region, to leave a desired array of dried,
relatively flat regions. This drying may be done by
heating or under vacuum.

In some cases, it is desired to first rehydrate
30   the droplets containing the analyte reagents to allow
for more time for adsorption to the solid support. It
is also possible to spot out the analyte reagents in a
humid environment so that droplets do not dry until the
arraying operation is complete.

19

III. Automated Apparatus for Forming Arrays

In another aspect, the invention includes an
automated apparatus for forming an array of analyte-
assay regions on a solid support, where each region in
5   the array has a known amount of a selected, analyte-
specific reagent.

The apparatus is shown in planar, and partially
schematic view in Fig. 4. A dispenser device 72 in the
apparatus has the basic construction described above
10   with respect to Fig. 1, and includes a dispenser 74
having an open-capillary channel terminating at a tip,
substantially as shown in Figs. 1 and 2A-2C.

The dispenser is mounted in the device for
movement toward and away from a dispensing position at
15   which the tip of the dispenser taps a support surface,
to dispense a selected volume of reagent solution, as
described above. This movement is effected by a
solenoid 76 as described above. Solenoid 76 is under
the control of a control unit 77 whose operation will
20   be described below. The solenoid is also referred to
herein as dispensing means for moving the device into
tapping engagement with a support, when the device is
positioned at a defined array position with respect to
that support.

25   The dispenser device is carried on an arm 74 which
is threadedly mounted on a worm screw 80 driven
(rotated) in a desired direction by a stepper motor 82
also under the control of unit 77. At its left end in
the figure screw 80 is carried in a sleeve 84 for
30   rotation about the screw axis. At its other end, the
screw is mounted to the drive shaft of the stepper
motor, which in turn is carried on a sleeve 86. The
dispenser device, worm screw, the two sleeves mounting
the worm screw, and the stepper motor used in moving
35   the device in the "x" (horizontal) direction in the

20

figure form what is referred to here collectively as a
displacement assembly 86.

The displacement assembly is constructed to
produce precise, micro-range movement in the direction

5   of the screw, i.e., along an x axis in the figure. In
one mode, the assembly functions to move the dispenser
in x-axis increments having a selected distance in the
range 5-25 μm. In another mode, the dispenser unit may
be moved in precise x-axis increments of several

10   microns or more,; for positioning the dispenser at
associated positions on adjacent supports, as will be
described below.

The displacement assembly, in turn, is mounted for
movement in the "y" (vertical) axis of the figure, for

15   positioning the dispenser at a selected y axis
position. The structure mounting the assembly includes
a fixed rod 88 mounted rigidly between a pair of frame
bars 90, 92, and a worm screw 94 mounted for rotation
between a pair of frame bars 96, 98. The worm screw is

20   driven (rotated) by a stepper motor 100 which operates
under the control of unit 77. The motor is mounted on
bar 96, as shown.

The structure just described, including worm screw
94 and motor 100, is constructed to produce precise,

25   micro-range movement in the direction of the screw,
i.e., along an y axis in the figure. As above, the
structure functions in one mode to move the dispenser
in y-axis increments having a selected distance in the
range 5-250 μm, and in a second mode, to move the

30   dispenser in precise y-axis increments of several
microns (μm) or more, for positioning the dispenser at
associated positions on adjacent supports.

The displacement assembly and structure for moving
this assembly in the y axis are referred to herein

35   collectively as positioning means for positioning the

21

dispensing device at a selected array position with respect to a support.

A holder 102 in the apparatus functions to hold a plurality of supports, such as supports 104 on which

5   the microarrays of regent regions are to be formed by the apparatus. The holder provides a number of recessed slots, such as slot 106, which receive the supports, and position them at precise selected positions with respect to the frame bars on which the

10   dispenser moving means is mounted.

As noted above, the control unit in the device functions to actuate the two stepper motors and dispenser solenoid in a sequence designed for automated operation of the apparatus in forming a selected

15   microarray of reagent regions on each of a plurality of supports.

The control unit is constructed, according to conventional microprocessor control principles, to provide appropriate signals to each of the solenoid and

20   each of the stepper motors, in a given timed sequence and for appropriate signalling time. The construction of the unit, and the settings that are selected by the user to achieve a desired array pattern, will be understood from the following description of a typical

25   apparatus operation.

Initially, one or more supports are placed in one or more slots in the holder. The dispenser is then moved to a position directly above a well (not shown) containing a solution of the first reagent to be

30   dispensed on the support(s). The dispenser solenoid is actuated now to lower the dispenser tip into this well, causing the capillary channel in the dispenser to fill. Motors 82, 100 are now actuated to position the dispenser at a selected array position at the first of

35   the supports. Solenoid actuation of the dispenser is

22

then effective to dispense a selected-volume droplet of
that reagent at this location. As noted above, this
operation is effective to dispense a selected volume
preferably between 2 pl and 2 nl of the reagent
5   solution.
        The dispenser is now moved to the corresponding
position at an adjacent support and a similar volume of
the solution is dispensed at this position. The
process is repeated until the reagent has been
10   dispensed at this preselected corresponding position on
each of the supports.
        Where it is desired to dispense a single reagent
at more than two array positions on a support, the
dispenser may be moved to different array positions at
15   each support, before moving the dispenser to a new
support, or solution can be dispensed at individual
positions on each support, at one selected position,
then the cycle repeated for each new array position.
        To dispense the next reagent, the dispenser is
20   positioned over a wash solution (not shown), and the
dispenser tip is dipped in and out of this solution
until the reagent solution has been substantially
washed from the tip. Solution can be removed from the
tip, after each dipping, by vacuum, compressed air
25   spray, sponge, or the like.
        The dispenser tip is now dipped in a second
reagent well, and the filled tip is moved to a second
selected array position in the first support. The
process of dispensing reagent at each of the
30   corresponding second-array positions is then carried as
above. This process is repeated until an entire
microarray of reagent solutions on each of the supports
has been formed.

35   IV. Microarray Substrate

23

This section describes embodiments of a substrate having a microarray of biological polymers carried on the substrate surface. Subsection A describes a multi-cell substrate, each cell of which contains a
5   microarray, and preferably an identical microarray, of distinct biopolymers, such as distinct polynucleotides, formed on a porous surface. Subsection B describes a microarray of distinct polynucleotides bound on a glass slide coated with a polycationic polymer.
10

     A.   Multi-Cell Substrate
     Fig. 9 illustrates, in plan view, a substrate 110 constructed according to the invention. The substrate has an 8 × 12 rectangular array 112 of cells, such as
15   cells 114, 116, formed on the substrate surface. With reference to Fig. 10, each cell, such as cell 114, in turn supports a microarray 118 of distinct biopolymers, such as polypeptides or polynucleotides at known, addressable regions of the microarray. Two such
20   regions forming the microarray are indicated at 120, and correspond to regions, such as regions 42, forming the microarray of distinct biopolymers shown in Fig. 3.
     The 96-cell array shown in Fig. 9 has typically array dimensions between about 12 and 244 mm in width
25   and 8 and 400 mm in length, with the cells in the array having width and length dimension of 1/12 and 1/8 the array width and length dimensions, respectively, *i.e.*, between about 1 and 20 in width and 1 and 50 mm in length.
30   The construction of substrate is shown cross-sectionally in Fig. 11, which is an enlarged sectional view taken along view line 124 in Fig. 9. The substrate includes a water-impermeable backing 126, such as a glass slide or rigid polymer sheet. Formed
35   on the surface of the backing is a water-permeable film

24

128. The film is formed of a porous membrane material, such as nitrocellulose membrane, or a porous web material, such as a nylon, polypropylene, or PVDF porous polymer material. The thickness of the film is

5   preferably between about 10 and 1000 μm. The film may be applied to the backing by spraying or coating uncured material on the backing, or by applying a preformed membrane to the backing. The backing and film may be obtained as a preformed unit from

10   commercial source, e.g., a plastic-backed nitrocellulose film available from Schleicher and Schuell Corporation.

With continued reference to Fig. 11, the film-covered surface in the substrate is partitioned into a

15   desired array of cells by water-impermeable grid lines, such as lines 130, 132, which have infiltrated the film down to the level of the backing, and extend above the surface of the film as shown, typically a distance of 100 to 2000 μm above the film surface.

20   The grid lines are formed on the substrate by laying down an uncured or otherwise flowable resin or elastomer solution in an array grid, allowing the material to infiltrate the porous film down to the backing, then curing or otherwise hardening the grid

25   lines to form the cell-array substrate.

One preferred material for the grid is a flowable silicone available from Loctite Corporation. The barrier material can be extruded through a narrow syringe (e.g., 22 gauge) using air pressure or

30   mechanical pressure. The syringe is moved relative to the solid support to print the barrier elements as a grid pattern. The extruded bead of silicone wicks into the pores of the solid support and cures to form a shallow waterproof barrier separating the regions of

35   the solid support.

25

In alternative embodiments, the barrier element can be a wax-based material or a thermoset material such as epoxy. The barrier material can also be a UV-curing polymer which is exposed to UV light after being

5 printed onto the solid support. The barrier material may also be applied to the solid support using printing techniques such as silk-screen printing. The barrier material may also be a heat-seal stamping of the porous solid support which seals its pores and forms a water-

10 impervious barrier element. The barrier material may also be a shallow grid which is laminated or otherwise adhered to the solid support.

In addition to plastic-backed nitrocellulose, the solid support can be virtually any porous membrane with

15 or without a non-porous backing. Such membranes are readily available from numerous vendors and are made from nylon, PVDF, polysulfone and the like. In an alternative embodiment, the barrier element may also be used to adhere the porous membrane to a non-porous

20 backing in addition to functioning as a barrier to prevent cross contamination of the assay reagents.

In an alternative embodiment, the solid support can be of a non-porous material. The barrier can be printed either before or after the microarray of

25 biomolecules is printed on the solid support.

As can be appreciated, the cells formed by the grid lines and the underlying backing are water-impermeable, having side barriers projecting above the porous film in the cells. Thus, defined-volume samples

30 can be placed in each well without risk of cross-contamination with sample material in adjacent cells. In Fig. 11, defined volumes samples, such as sample 134, are shown in the cells.

As noted above, each well contains a microarray of

35 distinct biopolymers. In one general embodiment, the

26

microarrays in the well are identical arrays of
distinct biopolymers, e.g., different sequence
polynucleotides. Such arrays can be formed in
accordance with the methods described in Section II, by
5   depositing a first selected polynucleotide at the same
selected microarray position in each of the cells, then
depositing a second polynucleotide at a different
microarray position in each well, and so on until a
complete, identical microarray is formed in each cell.
10   In a preferred embodiment, each microarray
contains about 10^3 distinct polynucleotide or
polypeptide biopolymers per surface area of less than
about 1 cm^2. Also in a preferred embodiment, the
biopolymers in each microarray region are present in a
15   defined amount between about 0.1 femtomoles and 100
nanomoles. The ability to form high-density arrays of
biopolymers, where each region is formed of a well-
defined amount of deposited material, can be achieved
in accordance with the microarray-forming method
20   described in Section II.
          Also in a preferred embodiments, the biopolymers
are polynucleotides having lengths of at least about 50
bp, i.e., substantially longer than oligonucleotides
which can be formed in high-density arrays by schemes
25   involving parallel, step-wise polymer synthesis on the
array surface.
          In the case of a polynucleotide array, in an assay
procedure, a small volume of the labeled DNA probe
mixture in a standard hybridization solution is loaded
30   onto each cell. The solution will spread to cover the
entire microarray and stop at the barrier elements.
The solid support is then incubated in a humid chamber
at the appropriate temperature as required by the
assay.

27

Each assay may be conducted in an "open-face" format where no further sealing step is required, since the hybridization solution will be kept properly hydrated by the water vapor in the humid chamber.   At
5   the conclusion of the incubation step, the entire solid support containing the numerous microarrays is rinsed quickly enough to dilute the assay reagents so that no significant cross contamination occurs. The entire solid support is then reacted with detection reagents.
10   if needed and analyzed using standard colorimetric, radioactive or fluorescent detection means. All . processing and detection steps are performed simultaneously to all of the microarrays on the solid support ensuring uniform assay conditions for all of
15   the microarrays on the solid support.

     B.   <u>Glass-Slide Polynucleotide Array</u>

          Fig. 5 shows a substrate 136 formed according to another aspect of the invention, and intended for use
20   in detecting binding of labeled polynucleotides to one or more of a plurality distinct polynucleotides. The substrate includes a glass substrate 138 having formed on its surface, a coating of a polycationic polymer, preferably a cationic polypeptide, such as polylysine
25   or polyarginine. Formed on the polycationic coating is a microarray 140 of distinct polynucleotides, each localized at known selected array regions, such as regions 142.
          The slide is coated by placing a uniform-thickness
30   film of a polycationic polymer, e.g., poly-l-lysine, on the surface of a slide and drying the film to form a dried coating. The amount of polycationic polymer added is sufficient to form at least a monolayer of polymers on the glass surface. The polymer film is
35   bound to surface via electrostatic binding between

28

negative silyl-OH groups n the surface and charged
amine groups in the polymers. Poly-1-lysine coated
glass slides may be obtained commercially, e.g., from
Sigma Chemical Co. (St. Louis, MO).

5   To form the microarray, defined volumes of
distinct polynucleotides are deposited on the polymer-
coated slide, as described in Section II. According to
an important feature of the substrate, the deposited
polynucleotides remain bound to the coated slide

10   surface non-covalently when an aqueous DNA sample is
applied to the substrate under conditions which allow
hybridization of reporter-labeled polynucleotides in
the sample to complementary-sequence (single-stranded)
polynucleotides in the substrate array. The method is

15   illustrated in Examples 1 and 2.

        To illustrate this feature, a substrate of the
type just described, but having an array of same-
sequence polynucleotides, was mixed with fluorescent-
labeled complementary DNA under hybridization

20   conditions. After washing to remove non-hybridized
material, the substrate was examined by low-power
fluorescence microscopy. The array can be visualized
by the relatively uniform labeling pattern of the array
regions.

25   In a preferred embodiment, each microarray
contains at least 10^3 distinct polynucleotide or
polypeptide biopolymers per surface area of less than
about 1 cm^2. In the embodiment shown in Fig. 5, the
microarray contains 400 regions in an area of about 16

30   mm^2, or 2.5 × 10^3 regions/cm^2. Also in a preferred
embodiment, the polynucleotides in the each microarray
region are present in a defined amount between about
0.1 femtomoles and 100 nanomoles in the case of
polynucleotides. As above, the ability to form high-

29

density arrays of this type, where each region is
formed of a well-defined amount of deposited material,
can be achieved in accordance with the microarray-
forming method described in Section II.

5   Also in a preferred embodiments, the
polynucleotides have lengths of at least about 50 bp,
i.e., substantially longer than oligonucleotides which
can be formed in high-density arrays by various in situ
synthesis schemes.

10

V.   Utility

Microarrays of immobilized nucleic acid sequences
prepared in accordance with the invention can be used
for large scale hybridization assays in numerous

15   genetic applications, including genetic and physical
mapping of genomes, monitoring of gene expression, DNA
sequencing, genetic diagnosis, genotyping of organisms,
and distribution of DNA reagents to researchers.

For gene mapping, a gene or a cloned DNA fragment

20   is hybridized to an ordered array of DNA fragments, and
the identity of the DNA elements applied to the array
is unambiguously established by the pixel or pattern of
pixels of the array that are detected. One application
of such arrays for creating a genetic map is described

25   by Nelson, et al. (1993). In constructing physical
maps of the genome, arrays of immobilized cloned DNA
fragments are hybridized with other cloned DNA
fragments to establish whether the cloned fragments in
the probe mixture overlap and are therefore contiguous

30   to the immobilized clones on the array. For example,
Lehrach, et al., describe such a process.

The arrays of immobilized DNA fragments may also
be used for genetic diagnostics. To illustrate, an
array containing multiple forms of a mutated gene or

35   genes can be probed with a labeled mixture of a

patient's DNA which will preferentially interact with only one of the immobilized versions of the gene.

The detection of this interaction can lead to a medical diagnosis. Arrays of immobilized DNA fragments

5   can also be used in DNA probe diagnostics. For example, the identity of a pathogenic microorganism can be established unambiguously by hybridizing a sample of the unknown pathogen's DNA to an array containing many types of known pathogenic DNA. A similar technique can

10   also be used for unambiguous genotyping of any organism. Other molecules of genetic interest, such as cDNA's and RNA's can be immobilized on the array or alternately used as the labeled probe mixture that is applied to the array.

15   In one application, an array of cDNA clones representing genes is hybridized with total cDNA from an organism to monitor gene expression for research or diagnostic purposes. Labeling total cDNA from a normal cell with one color fluorophore and total cDNA from a

20   diseased cell with another color fluorophore and simultaneously hybridizing the two cDNA samples to the same array of cDNA clones allows for differential gene expression to be measured as the ratio of the two fluorophore intensities. This two-color experiment can

25   be used to monitor gene expression in different tissue types, disease states, response to drugs, or response to environmental factors. & An example of this approach is illustrated in Examples 2, described with respect to Fig. 8.

30   By way of example and without implying a limitation of scope, such a procedure could be used to simultaneously screen many patients against all known mutations in a disease gene. This invention could be used in the form of, for example, 96 identical 0.9 cm ×

35   2.2 cm microarrays fabricated on a single 12 cm × 18 cm

31

sheet of plastic-backed nitrocellulose where each
microarray could contain, for example, 100 DNA
fragments representing all known mutations of a given
gene. The region of interest from each of the DNA
5   samples from 96 patients could be amplified, labeled,
and hybridized to the 96 individual arrays with each
assay performed in 100 microliters of hybridization
solution. The approximately 1 thick silicone rubber
barrier elements between individual arrays prevent
10   cross contamination of the patient samples by sealing
the pores of the nitrocellulose and by acting as a
physical barrier between each microarray. The solid
support containing all 96 microarrays assayed with the
96 patient samples is incubated, rinsed, detected and
15   analyzed as a single sheet of material using standard
radioactive, fluorescent, or colorimetric detection
means (Maniatas, et al., 1989). Previously, such a
procedure would involve the handling, processing and
tracking of 96 separate membranes in 96 separate sealed
20   chambers. By processing all 96 arrays as a single
sheet of material, significant time and cost savings
are possible.

The assay format can be reversed where the patient
or organism's DNA is immobilized as the array elements
25   and each array is hybridized with a different mutated
allele or genetic marker. The gridded solid support
can also be used for parallel non-DNA ELISA assays.
Furthermore, the invention allows for the use of all
standard detection methods without the need to remove
30   the shallow barrier elements to carry out the detection
step.

In addition to the genetic applications listed
above, arrays of whole cells, peptides, enzymes,
antibodies, antigens, receptors, ligands,
35   ph spholipids, polymers, drug cogener preparations or

32

chemical substances can be fabricated by the means
described in this invention for large scale screening
assays in medical diagnostics, drug discovery,
molecular biology, immunology and toxicology.

5   The multi-cell substrate aspect of the invention
allows for the rapid and convenient screening of many
DNA probes against many ordered arrays of DNA
fragments. This eliminates the need to handle and
detect many individual arrays for performing mass

10   screenings for genetic research and diagnostic
applications. Numerous microarrays can be fabricated
on the same solid support and each microarray reacted
with a different DNA probe while the solid support is
processed as a single sheet of material.

15

         The following examples illustrate, but in no way
are intended to limit, the present invention.

                           Example 1
20   Genomic-Complexity Hybridization to Micro
                DNA Arrays Representing the Yeast
                Saccharomyces cerevisiae Genome with
                   Two-Color Fluorescent Detection
         The array elements were randomly amplified PCR
25   (Bohlander, et al., 1992) products using physically
mapped lambda clones of S. cerevisiae genomic DNA
templates (Riles, et al., 1993). The PCR was performed
directly on the lambda phage lysates resulting in an
amplification of both the 35 kb lambda vector and the

30   5-15 kb yeast insert sequences in the form of a uniform
distribution of PCR product between 250-1500 base pairs
in length.   The PCR product was purified using
Sephadex G50 gel filtration (Pharmacia, Piscataway, NJ)
and concentrated by evaporation to dryness at room

35   temperature overnight.   Each of the 864 amplified

33

lambda clones was rehydrated in 15 µl of 3 × SSC in
preparation for spotting onto the glass.

    The micro arrays were fabricated on microscope
slides which were coated with a layer of poly-1-lysine
5   (Sigma). The automated apparatus described in Section
IV loaded 1 µl of the concentrated lambda clone PCR
product in 3 × SSC directly from 96 well storage plates
into the open capillary printing element and deposited
~5 nl of sample per slide at 380 micron spacing between
10   spots, on each of 40 slides. The process was repeated
for all 864 samples and 8 control spots. After the
spotting operation was complete, the slides were
rehydrated in a humid chamber for 2 hours, baked in a
dry 80° vacuum oven for 2 hours, rinsed to remove un-
15   absorbed DNA and then treated with succinic anhydride
to reduce non-specific adsorption of the labeled
hybridization probe to the poly-1-lysine coated glass
surface. Immediately prior to use, the immobilized DNA
on the array was denatured in distilled water at 90°
20   for 2 minutes.

    For the pooled chromosome experiment, the 16
chromosomes of *Saccharomyces cerevisiae* were separated
in a CHEF agarose gel apparatus (Biorad, Richmond, CA).
The six largest chromosomes were isolated in one gel
25   slice and the smallest 10 chromosomes in a second gel
slice. The DNA was recovered using a gel extraction
kit (Qiagen, Chatsworth, CA). The two chromosome pools
were randomly amplified in a manner similar to that
used for the target lambda clones. Following
30   amplification, 5 micrograms of each of the amplified
chromosome pools were separately random-primer labeled
using Klenow polymerase (Amersham, Arlington Heights,
IL) with a lissamine conjugated nucleotide analog
(Dupont NEN, Boston, MA) for the pool containing the
35   six largest chromosomes, and with a fluorescein

34

conjugated nucleotide analog (BMB) for the pool
containing smallest ten chromosomes. The two pools
were mixed and concentrated using an ultrafiltration
device (Amicon, Danvers, MA).

5   Five micrograms of the hybridization probe
consisting of both chromosome pools in 7.5 μl of TE was
denatured in a boiling water bath and then snap cooled
on ice. 2.5 μl of concentrated hybridization solution
(5 × SSC and 0.1% SDS) was added and all 10 μl
10   transferred to the array surface, covered with a cover
slip, placed in a custom-built single-slide humidity
chamber and incubated at 60° for 12 hours. The slides
were then rinsed at room temperature in 0.1 × SSC and
0.1%SDS for 5 minutes, cover slipped and scanned.

15   A custom built laser fluorescent scanner was used
to detect the two-color hybridization signals from the
1.8 × 1.8 cm array at 20 micron resolution. The
scanned image was gridded and analyzed using custom
image analysis software. After correcting for optical
20   crosstalk between the fluorophores due to their
overlapping emission spectra, the red and green
hybridization values for each clone on the array were
correlated to the known physical map position of the
clone resulting in a computer-generated color karyotype
25   of the yeast genome.

        Figure 6 shows the hybridization pattern of the
two chromosome pools. A red signal indicates that the
lambda clone on the array surface contains a cloned
genomic DNA segment from one of the largest six yeast
30   chromosomes. A green signal indicates that the lambda
clone insert comes from one of the smallest ten yeast
chromosomes. Orange signals indicate repetitive
sequences which cross hybridized to both chromosome
pools. Control spots on the array confirm that the
35   hybridization is specific and reproducible.

35

The physical map locations of the genomic DNA
fragments contained in each of the clones used as array
elements have been previously determined by Olson and
co-workers (Riles, *et al.*) allowing for the automatic
5   generation of the color karyotype shown in Figure 7.
The color of a chromosomal section on the karyotype
corresponds to the color of the array element
containing the clone from that section. The black
regions of the karyotype represent false negative dark
10   spots on the array (10%) or regions of the genome not
covered by the Olson clone library (90%). Note that
the largest six chromosomes are mainly red while the
smallest ten chromosomes are mainly green matching the
original CHEF gel isolation of the hybridization probe.
15   Areas of the red chromosomes containing green spots and
vice-versa are probably due to spurious sample tracking
errors in the formation of the original library and in
the amplification and spotting procedures.
The yeast genome arrays have also been probed with
20   individual clones or pools of clones that are
fluorescently labeled for physical mapping purposes.
The hybridization signals of these clones to the array
were translated into a position on the physical map of
yeast.
25

Example 2
Total cDNA Hybridized to Micro Arrays of
cDNA Clones with Two-Color
Fluorescent Detection
30   24 clones containing cDNA inserts from the plant
Arabidopsis were amplified using PCR. Salt was added
to the purified PCR products to a final concentration
of 3 × SSC. The cDNA clones were spotted on poly-1-
lysine coated microscope slides in a manner similar to
35   Example 1. Among the cDNA clones was a clone

36

representing a transcription factor HAT 4, which had
previously been used to create a transgenic line of the
plant *Arabidopsis*, in which this gene is present at ten
times the level found in wild-type *Arabidopsis* (Schena,
5   *et al*., 1992).

Total poly-A mRNA from wild type *Arabidopsis* was
isolated using standard methods (Maniatis, *et al*.,
1989) and reverse transcribed into total cDNA, using
fluorescein nucleotide analog to label the cDNA product
10   (green fluorescence).   A similar procedure was
performed with the transgenic line of *Arabidopsis* where
the transcription factor HAT4 was inserted into the
genome using standard gene transfer protocols.   cDNA
copies of mRNA from the transgenic plant are labeled
15   with a lissamine nucleotide analog (red fluorescence).
Two micrograms of the cDNA products from each type of
plant were pooled together and hybridized to the cDNA
clone array in a 10 microliter hybridization reaction
in a manner similar to Example 1.   Rinsing and
20   detection of hybridization was also performed in a
manner similar to Example 1.   Fig. 8 show the resulting
hybridization pattern of the array.

Genes equally expressed in wild type and the
transgenic *Arabidopsis* appeared yellow due to equal
25   contributions of the green and red fluorescence to the
final signal. The dots are different intensities of
yellow indicating various levels of gene expression.
The cDNA clone representing the transcription factor
HAT4, expressed in the transgenic line of *Arabidopsis*
30   but not detectably expressed in wild type *Arabidopsis*,
appears as a red dot (with the arrow pointing to it),
indicating the preferential expression of the
transcription factor in the red-labeled transgenic
Arabidopsis and the relative lack of expression of the

37

transcription factor in the green-labeled wild type
Arabidopsis.

An advantage of the microarray hybridization
format for gene expression studies is the high partial
5   concentration of each cDNA species achievable in the 10
microliter hybridization reaction. This high partial
concentration allows for detection of rare transcripts
without the need for PCR amplification of the
hybridization probe which may bias the true genetic
10   representation of each discrete cDNA species.

Gene expression studies such as these can be used
for genomics research to discover which genes are
expressed in which cell types, disease states,
development states or environmental conditions. Gene
15   expression studies can also be used for diagnosis of
disease by empirically correlating gene expression
patterns to disease states.

Example 3
Multiplexed Colorimetric Hybridization on
a Gridded Solid Support
20   A sheet of plastic-backed nitrocellulose was
gridded with barrier elements made from silicone rubber
according to the description in Section IV-A. The
sheet was soaked in 10 × SSC and allowed to dry. As
25   shown in Fig. 12, 192 M13 clones each with a different
yeast inserts were arrayed 400 microns apart in four
quadrants of the solid support using the automated
device described in Section III. The bottom left
30   quadrant served as a negative control for hybridization
while each of the other three quadrants was hybridized
simultaneously with a different oligonucleotide using
the open-face hybridization technology described in
Section IV-A. The first two and last four elements of

38

each array are positive controls for the colorimetric
detection step.

The oligonucleotides were labeled with fluorescein
which was detected using an anti-fluorescein antibody
5   conjugated to alkaline phosphatase that precipitated an
NBT/BCIP dye on the solid support (Amersham). Perfect
matches between the labeled oligos and the M13 clones
resulted in dark spots visible to the naked eye and
detected using an optical scanner (HP ScanJet II)
10   attached to a personal computer. The hybridization
patterns are different in every quadrant indicating
that each oligo found several unique M13 clones from
among the 192 with a perfect sequence match. Note that
the open capillary printing tip leaves detectable
15   dimples on the nitrocellulose which can be used to
automatically align and analyze the images.

Although the invention has been described with
respect to specific embodiments and methods, it will be
20   clear that various changes and modification may be made
without departing from the invention.

39

IT IS CLAIMED:

1.   A method of forming a microarray of analyte-
assay regions on a solid support, where each region in
5   the array has a known amount of a selected, analyte-
specific reagent, said method comprising, ·

(a) loading a solution of a selected analyte-
specific reagent in a reagent-dispensing device having
an elongate capillary channel (i) formed by spaced-
10   apart, coextensive elongate members, (ii) adapted to
hold a quantity of the reagent solution and (iii)
having a tip region at which aqueous solution in the
channel forms a meniscus,

(b) tapping the tip of the dispensing device
15   against a solid support at a defined position on the
surface, with an impulse effective to break the
meniscus in the capillary channel and deposit a
selected volume of solution on the surface, and ·

(c) repeating steps (a) and (b) until said array
20   is formed.

2.   The method of claim 1, wherein said tapping is
carried out with an impulse effective to deposit a
selected volume in the volume range between 0.01 to 100
25   nl.

3.   The method of claim 1, wherein said channel is
formed by a pair of spaced-apart tapered elements.

30   4.   The method of claim 1, for forming a plurality
of such arrays, wherein step (b) is applied to a
selected position on each of a plurality of solid
supports at each repeat cycle proceeding step (c).

40

5.   The method of claim 1, which further includes, after performing steps (a) and (b) at least one time, reloading the reagent-dispensing device with a new reagent solution by the steps of (i) dipping the
5   capillary channel of the device in a wash solution, (ii) removing wash solution drawn into the capillary channel, and (iii) dipping the capillary channel into the new reagent solution.

10   6.   Automated apparatus for forming a microarray of analyte-assay regions on a plurality of solid supports, where each region in the array has a known amount of a selected, analyte-specific reagent, said apparatus comprising
15   (a) a holder for holding, at known positions, a plurality of planar supports,
          (b) a reagent dispensing device having an open capillary channel (i) formed by spaced-apart, coextensive elongate members (ii) adapted to hold a
20   quantity of the reagent solution and (iii) having a tip region at which aqueous solution in the channel forms a meniscus,
          (c) positioning means for positioning the dispensing device at a selected array position with
25   respect to a support in said holder,
          (d) dispensing means for moving the device into tapping engagement against a support with a selected impulse, when the device is positioned at a defined array position with respect to that support, with an
30   impulse effective to break the meniscus of liquid in the capillary channel and deposit a selected volume of solution on the surface, and
          (e) control means for controlling said positioning and dispensing means.

35

41

7.   The apparatus of claim 6, wherein said
dispensing means is effective to move said dispensing
device against a support with an impulse effective to
deposit a selected volume in the volume range between
5   0.01 to 100 nl.

8.   The apparatus of claim 6, wherein said channel
is formed by a pair of spaced-apart tapered elements.

10   9.   The apparatus of claim 6, wherein the control
means operates to (i) place the dispensing device at a
loading station, (ii) move the capillary channel in the
device into a selected reagent at the loading station,
to load the dispensing device with the reagent, and
15   (iii) dispense the reagent at a defined array position
on each of the supports on said holder.

10.   The apparatus of claim 6, wherein the control
device further operates, at the end of a dispensing
20   cycle, to wash the dispensing device by (i) placing the
dispensing device at a washing station, (ii) moving the
capillary channel in the device into a wash fluid, to
load the dispensing device with the fluid, and (iii)
remove the wash fluid prior to loading the dispensing
25   device with a fresh selected reagent.

11.   The apparatus of claim 6, wherein said device
is one of a plurality of such devices which are carried
on the arm for dispensing different analyte assay
30   reagents at selected spaced array positions.

12.   A substrate with a surface having a
microarray of at least 10^3 distinct polynucleotide or
polypeptide biopolymers per 1 cm^2 surface area, each

42

distinct biopolymer sample (i) being disposed at a
separate, defined position in said array, (ii) having a
length of at least 50 subunits, and (iii) being present
in a defined amount between about 0.1 femtomole and 100
5   nanomoles.

13.   The substrate of claim 12, wherein said
surface is glass slide coated with polylysine, and said
biopolymers are polynucleotides.
10

14.   The substrate of claim 12, wherein said
substrate has a water-impermeable backing, a water-
permeable film formed on the backing, and a grid formed
on the film, where said grid (i) is composed of
15   intersecting water-impervious grid elements extending
from said backing to positions raised above the surface
of said film, and (ii) partitions the film into a
plurality of water-impervious cells, where each cell
contains such a biopolymer array.
20

15.   A substrate with a surface array of sample-
receiving cells, comprising
       a water-impermeable backing,
       a water-permeable film formed on the backing, and
25   a grid formed on the film, said grid being composed of
intersecting water-impervious grid elements extending
from said backing to positions raised above the surface
of said film.

30   16.   The substrate of claim 15, wherein the cells
of the array each contain an array of biopolymers.

17.   A substrate for use in detecting binding of
labeled biopolymers to one or more of a plurality
35   distinct polynucleotides, comprising

43

a non-porous, glass substrate,

a coating of a cationic polymer on said substrate,
and

an array of distinct polynucleotides to said
5   coating, where each biopolymer is disposed at a
separate, defined position in a surface array of
biopolymers.

18.   A method of detecting differential expression
10   of each of a plurality of genes in a first cell type
with respect to expression of the same genes in a
second cell types, said method comprising

producing fluorescence-labeled cDNA's from mRNA's
isolated from the two cells types, where the cDNA's
15   from the first and second cells are labeled with first
and second different fluorescent reporters,

adding a mixture of the labeled cDNA's from the
two cell types to an array of polynucleotides
representing a plurality of known genes derived from
20   the two cell types, under conditions that result in
hybridization of the cDNA's to complementary-sequence
polynucleotides in the array; and

examining the array by fluorescence under
fluorescence excitation conditions in which (i)
25   polynucleotides in the array that are hybridized
predominantly to cDNA's derived from one of the first
and second cell types give a distinct first or second
fluorescence emission color, respectively, and (ii)
polynucleotides in the array that are hybridized to
30   substantially equal numbers of cDNA's derived from the
first and second cell types give a distinct combined
fluorescence emission color, respectively,

wherein the relative expression of known genes in
the two cell types can be determined by the observed
35   fluorescence emission color of each spot.

44

19. The method of claim 18, wherein the array of
polynucleotides is formed on a substrate with a surface
having an array of at least 10^2 distinct polynucleotide
or polypeptide biopolymers in a surface area of less

5   than about 1 cm², each distinct biopolymer (i) being
disposed at a separate, defined position in said array,
(ii) having a length of at least 50 subunits, and (iii)
being present in a defined amount between about .1
femtomole and 100 nmoles.

10

20. The method of claim 19, wherein said surface
is a glass slide coated with polylysine, and said
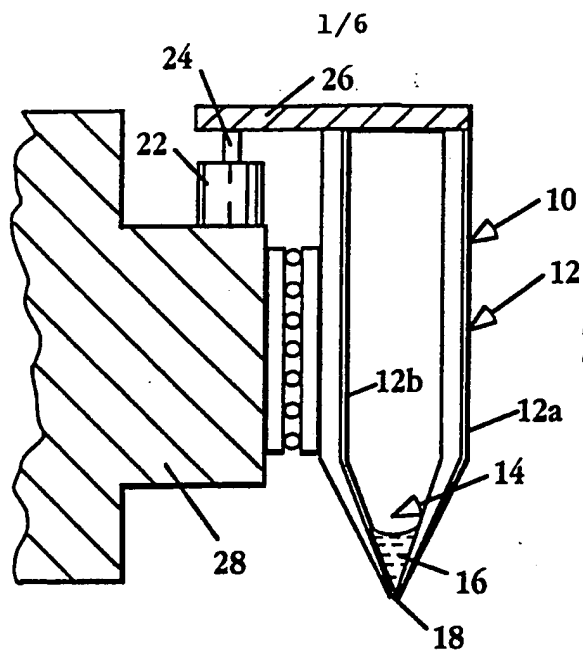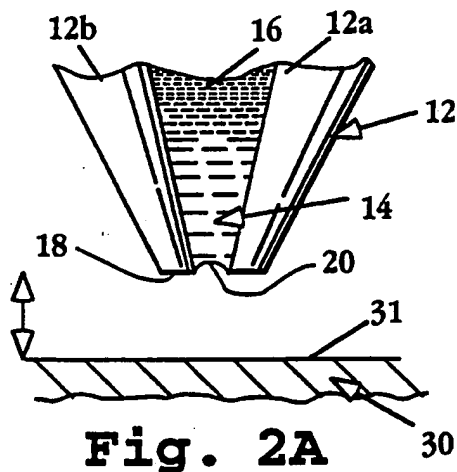biopolymers are polynucleotides non-covalently bound to
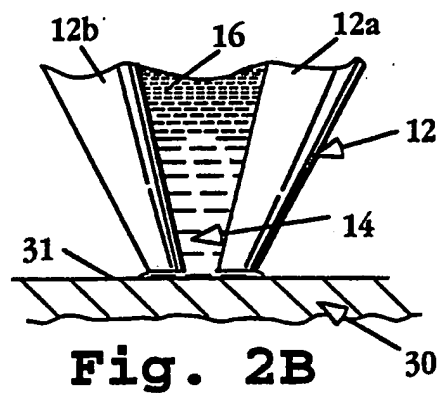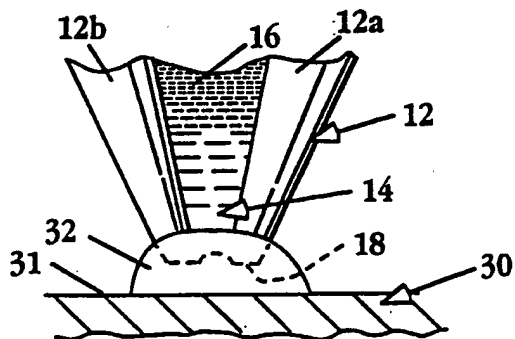said polylysine.

15

Fig. 1

Fig. 2A

Fig. 2B

Fig. 2C

Fig. 3

Fig. 4

Fig. 5

Fig. 6

Fig. 7

Fig. 8

Fig. 9

Fig. 10

Fig. 11

Fig. 12

INTERNATIONAL SEARCH REPORT

A. CLASSIFICATION OF SUBJECT MATTER
IPC(6) :G01N 33/543, 33/68
US CL :435/6; 436/518
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED
Minimum documentation searched (classification system followed by classification symbols)
U.S. : 422/57; 435/4,6,973; 436/518,524,527,531,805,809

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

C. DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A,P	US, A, 5,338,688 (DEEG ET AL) 16 August 1994, see entire document	1-17
A	US, A, 5,204,268 (MATSUMOTO) 20 April 1993, see entire document.	6-11
A	US, A, 4,071,315 (CHATEAU) 31 January 1978, see entire document.	12-17
A	US, A, 5,100,777 (CHANG) 31 March 1992, see entire document.	12-17
A	US, A, 5,200,312 (OPRANDY) 06 April 1993, see entire document.	12-17

☐ Further documents are listed in the continuation of Box C.   ☐ See patent family annex.

*	Special categories of cited documents:
"A"	document defining the general state of the art which is not considered to be of particular relevance
"E"	earlier document published on or after the international filing date
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O"	document referring to an oral disclosure, use, exhibition or other means
"P"	document published prior to the international filing date but later than the priority date claimed

"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
15 SEPTEMBER 1995	06 OCT 1995
Name and mailing address of the ISA/US Commissioner of Patents and Trademarks Box PCT Washington. D.C. 20231	Authorized officer CHRISTOPHER CHIN
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

Discovery and analysis of inflammatory disease-related genes using cDNA microarrays

(inflammation/human genome analysis/gene discovery)

RENU A. HELLER*†, MARK SCHENA*, ANDREW CHAI*, DARI SHALON‡, TOD BEDILION‡, JAMES GILMORE‡, DAVID E. WOOLLEY§, AND RONALD W. DAVIS*

*Department of Biochemistry, Beckman Center, Stanford University Medical Center, Stanford, CA 94305; ‡Synteni, Palo Alto, CA 94306; and §Department of Medicine, Manchester Royal Infirmary, Manchester, United Kingdom

Contributed by Ronald W. Davis, December 27, 1996

ABSTRACT   cDNA microarray technology is used to profile complex diseases and discover novel disease-related genes. In inflammatory disease such as rheumatoid arthritis, expression patterns of diverse cell types contribute to the pathology. We have monitored gene expression in this disease state with a microarray of selected human genes of probable significance in inflammation as well as with genes expressed in peripheral human blood cells. Messenger RNA from cultured macrophages, chondrocyte cell lines, primary chondrocytes, and synoviocytes provided expression profiles for the selected cytokines, chemokines, DNA binding proteins, and matrix-degrading metalloproteinases. Comparisons between tissue samples of rheumatoid arthritis and inflammatory bowel disease verified the involvement of many genes and revealed novel participation of the cytokine interleukin 3, chemokine Groα and the metalloproteinase matrix metallo-elastase in both diseases. From the peripheral blood library, tissue inhibitor of metalloproteinase 1, ferritin light chain, and manganese superoxide dismutase genes were identified as expressed differentially in rheumatoid arthritis compared with inflammatory bowel disease. These results successfully demonstrate the use of the cDNA microarray system as a general approach for dissecting human diseases.

The recently described cDNA microarray or DNA-chip technology allows expression monitoring of hundreds and thousands of genes simultaneously and provides a format for identifying genes as well as changes in their activity (1, 2). Using this technology, two-color fluorescence patterns of differential gene expression in the root versus the shoot tissue of *Arabidopsis* were obtained in a specific array of 48 genes (1). In another study using a 1000 gene array from a human peripheral blood library, novel genes expressed by T cells were identified upon heat shock and protein kinase C activation (3).

The technology uses cDNA sequences or cDNA inserts of a library for PCR amplification that are arrayed on a glass slide with high speed robotics at a density of 1000 cDNA sequences per cm². These microarrays serve as gene targets for hybridization to cDNA probes prepared from RNA samples of cells or tissues. A two-color fluorescence labeling technique is used in the preparation of the cDNA probes such that a simultaneous hybridization but separate detection of signals provides the comparative analysis and the relative abundance of specific genes expressed (1, 2). Microarrays can be constructed from specific cDNA clones of interest, a cDNA library, or a select number of open reading frames from a genome sequencing database to allow a large-scale functional analysis of expressed sequences.

Because of the wide spectrum of genes and endogenous mediators involved, the microarray technology is well suited for analyzing chronic diseases. In rheumatoid arthritis (RA), inflammation of the joint is caused by the gene products of many different cell types present in the synovium and cartilage tissues plus those infiltrating from the circulating blood. The autoimmune and inflammatory nature of the disease is a cumulative result of genetic susceptibility factors and multiple responses, paracrine and autocrine in nature, from macrophages, T cells, plasma cells, neutrophils, synovial fibroblasts, chondrocytes, etc. Growth factors, inflammatory cytokines (4), and the chemokines (5) are the important mediators of this inflammatory process. The ensuing destruction of the cartilage and bone by the invading synovial tissue includes the actions of prostaglandins and leukotrienes (6), and the matrix degrading metalloproteinases (MMPs). The MMPs are an important class of Zn-dependent metallo-endoproteinases that can collectively degrade the proteoglycan and collagen components of the connective tissue matrix (7).

This paper presents a study in which the involvement of select classes of molecules in RA was examined. Also investigated were 1000 human genes randomly selected from a peripheral human blood cell library. Their differential and quantitative expression analysis in cells of the joint tissue, in diseased RA tissue and in inflammatory bowel disease (IBD) tissues was conducted to demonstrate the utility of the microarray method to analyze complex diseases by their pattern of gene expression. Such a survey provides insight not only into the underlying cause of the pathology, but also provides the opportunity to selectively target genes for disease intervention by appropriate drug development and gene therapies.

METHODS

Microarray Design, Development, and Preparation. Two approaches for the fabrication of cDNA microarrays were used in this study. In the first approach, known human genes of probable significance in RA were identified. Regions of the clones, preferably 1 kb in length, were selected by their proximity to the 3' end of the cDNA and for areas of least identity to related and repetitive sequences. Primers were synthesized to amplify the target regions by standard PCR protocols (3). Products were

Biochemistry: Heller *et al.*

Proc. Natl. Acad. Sci. USA 94 (1997)   2151

verified by gel electrophoresis and purified with Qiaquick 96-well purification kit (Qiagen, Chatsworth, CA), lyophilized (Savant), and resuspended in 5 μl of 3× standard saline citrate (SSC) buffer for arraying. In the second approach, the microarray containing the 1056 human genes from the peripheral blood lymphocyte library was prepared as described (3).

Tissue Specimens. Rheumatoid synovial tissue was obtained from patients with late stage classic RA undergoing remedial synovectomy or arthroplasty of the knee. Synovial tissue was separated from any associated connective tissue or fat. One gram of each synovial specimen was subjected to RNA extraction within 40 min of surgical excision, or explants were cultured in serum-free medium to examine any changes under *in vitro* conditions. For IBD, specimens of macroscopically inflamed lower intestinal mucosa were obtained from patients with Crohn disease undergoing remedial surgery. The hypertrophied mucosal tissue was separated from underlying connective tissue and extracted for RNA.

Cultured Cells. The Mono Mac-6 (MM6) monocytic cells (8) were grown in RPMI medium. Human chondrosarcoma SW1353 cells, primary human chondrocytes, and synoviocytes (9, 10) were cultured in DMEM; all culture media were supplemented with 10% fetal bovine serum, 100 μg/ml streptomycin, and 500 units/ml penicillin. Treatment of cells with lipopolysaccharide (LPS) endotoxin at 30 ng/ml, phorbol 12-myristate 13-acetate (PMA) at 50 ng/ml, tumor necrosis factor α (TNF-α) at 50 ng/ml, interleukin (IL)-1β at 30 ng/ml, or transforming growth factor-β (TGF-β) at 100 ng/ml is described in the figure legends.

Fluorescent Probe, Hybridization, and Scanning. Isolation of mRNA, probe preparation, and quantitation with *Arabidopsis* control mRNAs was essentially as described (3) except for the following minor modification. Following the reverse transcriptase step, the appropriate Cy3- and Cy5-labeled samples were pooled; mRNA degraded by heating the sample to 65°C for 10 min with the addition of 5 μl of 0.5M NaOH plus 0.5 ml of 10 mM EDTA. The pooled cDNA was purified from unincorporated nucleotides by gel filtration in Centri-spin columns (Princeton Separations, Adelphia, NJ). Samples were lyophilized and dissolved in 6 μl of hybridization buffer (5× SSC plus 0.2% SDS). Hybridizations, washes, scanning, quantitation procedures, and pseudocolor representations of fluorescent images have been described (3). Scans for the two fluorescent probes were normalized either to the fluorescence intensity of *Arabidopsis* mRNAs spiked into the labeling reactions (see Figs. 2–4) or to the signal intensity of β-actin and glyceraldehyde-3-phosphate dehydrogenase (GAPDH; see Fig. 5).

RESULTS

Ninety-Six-Gene Microarray Design. The actions of cytokines, growth factors, chemokines, transcription factors, MMPs, prostaglandins, and leukotrienes are well recognized in inflammatory disease, particularly RA (11–14). Fig. 1 displays the selected genes for this study and also includes control cDNAs of housekeeping genes such as β-actin and GAPDH and genes from *Arabidopsis* for signal normalization and quantitation (row A, columns 1–12).

Defining Microarray Assay Conditions. Different lengths and concentrations of target DNA were tested by arraying PCR-

	1	2	3	4	5	6	7	8	9	10	11	12
A	BLANK	BLANK	HAT1 / HAT1	HAT1 / HAT1	HAT4 / HAT4	HAT4 / HAT4	HAT22 / HAT22	HAT22 / HAT22	YES23 / YES23	YES23 / YES23	BACTIN / β-actin	G3PDH / G3PDH
B	IL1A / IL-1α	IL1B / IL-1β	IL1RA / IL-1RA	IL2 / IL-2	IL3 / IL-3	IL4 / IL-4	IL6 / IL-6	IL6R / IL-6R	IL7 / IL-7	CFOS / c-fos	CJUN / c-jun	RFRA1 / Rat Fra-1
C	IL8 / IL-8	IL9 / IL-9	IL10 / IL-10	ICE / ICE	IFNG / IFNγ	GCSF / G-CSF	MCSF / M-CSF	GMCSF / GM-CSF	TNFB.1 / TNFβ	CREL / c-rel	NFKB50 / NFκBp50	NFKB65.1 / NFκBp65
D	TNFA.1 / TNFα	TNFA.2 / TNFα	TNFA.3 / TNFα	TNFA.4 / TNFα	TNFA.5 / TNFα	TNFRI.1 / TNFrI	TNFRI.2 / TNFrI	TNFRII.1 / TNFrII	TNFRII.2 / TNFrII	NFKB65.2 / NFκBp65	IKB / IκB	CREB2 / CREB2
E	STR1 / Strom-1	STR2-3 / Strom-2	STR3 / Strom-3	COL1 / Coll-1	COL1-3' / Coll-1.3'	COL2.1 / Coll-2	COL2.2 / Coll-2	COL3 / Coll-3	COX1 / Cox-1	COX2 / Cox-2	12LO / 12-LO	15LO / 15-LO
F	GELA.1 / Gel-A	GELB / Gel-B	HME / Elastase	MTMMP / MT-MMP	PUMP1 / Matrilysin	TIMP1 / TIMP-1	TIMP2 / TIMP-2	TIMP3 / TIMP-3	ICAM1 / ICAM-1	VCAM / VCAM	5LO.1 / 5-LO	CPLA2.2 / cPLA2
G	EGF / EGF	FGFA / FGF acidic	FGFB / FGF basic	IGFI / IGF-I	IGFII / IGF-II	TGFA / TGFα	TGFB / TGFβ	PDGFB / PDGFβ	CALCTN / Calcitonin	GH1 / GH-1	GRO / GRO1α	GCR / GR
H	MCP1.1 / MCP-1	MCP1.1 / MCP-1	MIP1A / MIP-1α	MIP1B / MIP-1β	MIF / MIF	RANTES / RANTES	INOS / iNOS	LDLR / LDLR	ALU.1 / IL-10	ALU.2 / TNFRp70	ALU.3 / IL-10	POLYA / LDLR

- ▧ *A. thaliana* controls
- ▨ Human controls
- ☐ Cytokines and related genes
- ☐ Transcription factors and related genes
- ☐ MMP's and related genes
- ☐ Chemokines
- ☐ Growth factors and related genes
- ☐ Other genes

Fig. 1. Ninety-six-element microarray design. The target element name and the corresponding gene are shown in the layout. Some genes have more than one target element to guarantee specificity of signal. For TNF the targets represent decreasing lengths of 1, 0.8, 0.6, 0.4, and 0.2 kb from left to right.

amplified products ranging from 0.2 to 1.2 kb at concentrations of 1 μg/μl or less. No significant difference in the signal levels was observed within this range of target size and only with 0.2-kb length was a signal reduced upon an 8-fold dilution of the 1 μg/μl sample (data not shown). In this study the average length of the targets was 1 kb, with a few exceptions in the range of ≈300 bp, arrayed at a concentration of 1 μg/μl. Normally one PCR provided sufficient material to fabricate up to 1000 microarray targets.

In considering positional effects in the development of the targets for the microarrays, selection was biased toward the 3′ proximal regions, because the signal was reduced if the target fragment was biased toward the 5′ end (data not shown). This result was anticipated since the hybridizing probe is prepared by reverse transcription with oligo(dT)-primed mRNA and is richer in 3′ proximal sequences. Cross-hybridizations of probes to targets of a gene family were analyzed with the matrix metal-

loproteinases as the example because they can show regions of sequence identities of greater than 70%. With collagenase-1 (Col-1) and collagenase-2 (Col-2) genes as targets with up to 70% sequence identity, and stromelysin-1 (Strom-1) and stromelysin-2 (Strom-2) genes with different degrees of identity, our results showed that a short region of overlap, even with 70–90% sequence identity, produced a low level of cross-hybridization. However, shorter regions of identity spread over the length of the target resulted in cross-hybridization (data not shown). For closely related genes, targets were designed by avoiding long stretches of homology. For members of a gene family two or more target regions were included to discriminate between specificity of signal versus cross-hybridization.

Monitoring Differential Expression in Cultured Cell Lines. In RA tissue, the monocyte/macrophage population plays a prominent role in phagocytic and immunomodulatory activities. Typ-

FIG. 2.   Time course for LPS/PMA-induced MM6 cells. Array elements are described in Fig. 1. (*A*) Pseudocolor representations of fluorescent scans correspond to gene expression levels at each time point. The array is made up of 8 *Arabidopsis* control targets and 86 human cDNA targets, the majority of which are genes with known or suspected involvement in inflammation. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation. Fluorescent probes were made by labeling mRNA from untreated MM6 cells or LPS and PMA treated cells. mRNA was isolated at indicated times after induction. (*B I–III*) The two-color samples were cohybridized, and microarray scans provided the data for the levels of select transcripts at different time points relative to abundance at time zero. The analysis was performed using normalized data collected from 8-bit images.

Biochemistry: Heller *et al.*

Proc. Natl. Acad. Sci. USA 94 (1997)   2153

ically these cells, when triggered by an immunogen, produce the proinflammatroy cytokines TNF and IL-1. We have used the monocyte cell line MM6 and monitored changes in gene expression upon activation with LPS endotoxin, a component of Gram-negative bacterial membranes, and PMA, which augments the action of LPS on TNF production (15). RNA was isolated at different times after induction and used for cDNA probe preparation. From this time course it was clear that TNF expression was induced within 15 min of treatment, reached maximum levels in 1 hr, remained high until 4 hr and subsequently declined (Fig. 2*A*). Many other cytokine genes were also transiently activated, such as IL-1α and -β, IL-6, and granulocyte colony-stimulating factor (GCSF). Prominent chemokines activated were IL-8, macrophage inflammatory protein (MIP)-1β, more so than MIP-1α, and Groα or melanoma growth stimulatory factor. Migration inhibitory factor (MIF) expressed in the uninduced state declined in LPS-activated cells. Of the immediate early genes, the noticeable ones were c-*fos*, *fra-1*, c-*jun*, NF-κBp50, and IκB, with c-*rel* expression observed even in the uninduced state (Fig. 2*B*). These expression patterns are consistent with reported patterns of activation of certain LPS- and PMA-induced genes (12). Demonstrated here is the unique ability of this system to allow parallel visualization of a large number of gene activities over a period of time.

SW1353 cells is a line derived from malignant tumors of the cartilage and behaves much like the chondrocytes upon stimulation with TNF and IL-1 in the expression of MMPs (9). In addition to confirming our earlier observations with Northern blots on Strom-1, Col-1, and Col-3 expression (9), gelatinase (Gel) A, putative metalloproteinase (PUMP)-1 membrane-

type matrix metalloproteinase, tissue inhibitors of matrix metalloproteinases or tissue inhibitor of metalloproteinase 1 (TIMP-1), -2, and -3 were also expressed by these cells together with the human matrix metallo-elastase (HME; Fig. 3*A*). HME induction was estimated to be ≈50-fold and was greater than any of the other MMPs examined (Fig. 3*B*). This result was unexpected because HME is reportedly expressed only by alveolar macrophage and placental cells (16). Expression of the cytokines and chemokines, IL-6, IL-8, MIF, and MIP-1β was also noted. A variety of other genes, including certain transcription factors, were also up-regulated (Fig. 3), but the overall time-dependent expression of genes in the SW1353 cells was qualitatively distinct from the MM6 cells.

Quantitation of differential gene expression (Figs. 2*B* and 3*B*) was achieved with the simultaneous hybridization of Cy3-labeled cDNA from untreated cells and Cy5-labeled cDNA from treated samples. The estimated increases in expression from these microarrays for a select number of genes including IL-1β, IL-8, MIP-1β, TNF, HME, Col-1, Col-3, Strom-1, and Strom-2 were compared with data collected from dot blot analysis. Results (not shown) were in close agreement and confirmed our earlier observations on the use of the microarray method for the quantitation of gene expression (3).

Expression Profiles in Primary Chondrocytes and Synoviocytes of Human RA Tissue. Given the sensitivity and the specificity of this method, expression profiles of primary synoviocytes and chondrocytes from diseased tissue were examined. Without prior exposure to inducing agents, low level expression of c-*jun*, GCSF, IL-3, TNF-β, MIF, and RANTES (regulated upon activation, normal T cell expressed and secreted) was seen as well as expression of MMPs, GelA, Strom-1, Col-1, and the three TIMPs. In this case, Col-2 hybridization was considered to be nonspecific because the second Col-2 target taken from the 3' end of the gene gave no

FIG. 3. Time course for IL-1β and TNF-induced SW1353 cells using the inflammation array (Fig. 1). (*A*) Pseudocolor representation of fluorescent scans correspond to gene expression levels at each time point. (*B I–IV*) Relative levels of selected genes at different time points compared with time zero.

FIG. 4. Expression profiles for early passage primary synoviocytes and chondrocytes isolated from RA tissue, cultured in the presence of 10% fetal calf serum and activated with PMA and IL-1β, or TNF and IL-1β, or TGF-β for 18 hr. The color bars provide a comparative calibration scale between arrays and are derived from the *Arabidopsis* mRNA samples that are introduced in equal amounts during probe preparation

signal. Treatment more so with PMA and IL-1, than TNF and IL-1, produced a dramatic up-regulation in expression of several genes in both of these primary cell types. These genes are as follows: the cytokine IL-6, the chemokines IL-8 and Gro-1α, and the MMPs; Strom-1, Col-1, Col-3, and HME; and the adhesion molecule, vascular cell adhesion molecule 1 (VCAM-1). The surprise again is HME expression in these primary cells, for reasons discussed above. From these results, the expression profiles of synoviocytes and the chondrocytes appear very similar; the differences are more quantitative than qualitative. Treatment of the primary chondrocytes with the anabolic growth factor TGF-β had an interesting profile in that it produced a remarkable down-regulation of genes expressed in both the untreated and induced state (Fig. 4).

Given the demonstrated effectiveness of this technology, a comparative analysis of two different inflammatory disease states was conducted with probes made from RA tissue and IBD samples. RA samples were from late stage rheumatoid synovial tissue, and IBD specimens were obtained from inflamed lower intestinal mucosa of patients with Crohn disease. With both the 96-element known gene microarray and the 1000-gene microarray of cDNAs selected from a peripheral human blood cell library (3), distinct differences in gene expression patterns were evident. On the 96-gene array, RA tissue samples from different affected individuals gave similar profiles (data not shown) as did different samples from the same individual (Fig. 5). These patterns were notably similar to those observed with primary synoviocytes and chondrocytes (Fig. 4). Included in the list of prominently up-regulated genes are IL-6, the MMPs Strom-1, Col-1, GelA, HME, and in

certain samples PUMP, TIMPs, particularly TIMP-1 and TIMP-3, and the adhesion molecule VCAM. Discernible levels of macrophage chemotactic protein 1 (MCP-1), MIF and RANTES were also noted. IBD samples were in comparison, rather subdued although IL-1 converting enzyme (ICE), TIMP-1, and MIF were notable in all the three different IBD samples examined here. In IBD-A, one of three individual samples, ICE, VCAM, Groα, and MMP expression was more pronounced than in the others.

We also made use of a peripheral blood cDNA library (3) to identify genes expressed by lymphocytes infiltrating the inflamed tissues from the circulating blood. With the 1046-element array of randomly selected cDNAs from this library, probes made from RA and IBD samples showed hybridizations to a large number of genes. Of these, many were common between the two disease tissues while others were differentially expressed (data not shown). A complete survey of these genes was beyond the scope of this study, but for this report we picked three genes that were up-regulated in the RA tissue relative to IBD. These cDNAs wer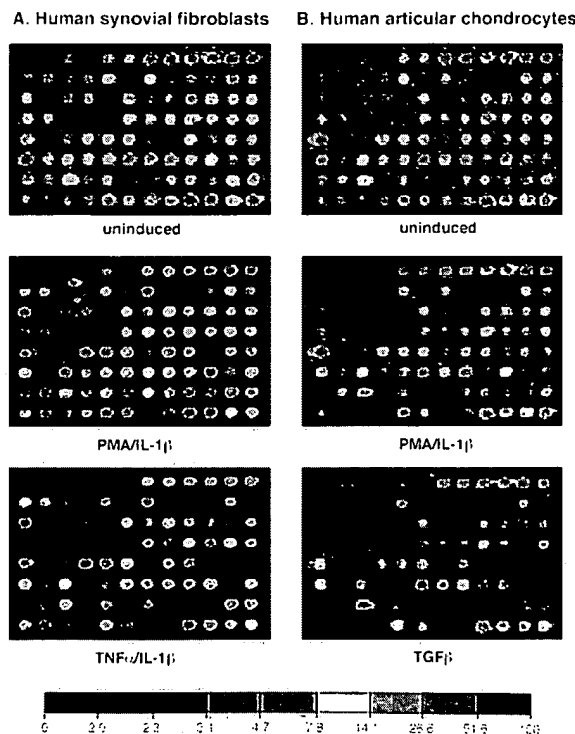e sequenced and identified by comparison to the GenBank database. They are TIMP-1, apoferritin light chain, and manganese superoxide dismutase (MnSOD). Differential expression of MnSOD was only observed in samples of RA tissue explants maintained in growth medium without serum for anywhere between 2 to 16 hr. These results also indicate that the expression profile of genes can be altered when explants are transferred to culture conditions.

DISCUSSION

The speed, ease, and feasibility of simultaneously monitoring differential expression of hundreds of genes with the cDNA microarray based system (1–3) is demonstrated here in the analysis of a complex disease such as RA. Many different cell types in the RA tissue; macrophages, lymphocytes, plasma cells, neutrophils, synoviocytes, chondrocytes, etc. are known to contribute to the development of the disease with the expression of gene products known to be proinflammatory. They include the cytokines, chemokines, growth factors, MMPs, eicosanoids, and others (7, 11–14), and the design of the 96-element known gene microarray was based on this knowledge and depended on the availability of the genes. The technology was validated by confirming earlier observations on the expression of TNF by the monocyte cell line MM6, and of Col-1 and Col-3 expression in the chondrosarcoma cells and articular chondrocytes (9, 12). In our time-dependent survey the chronological order of gene activities in and between gene families was compared and the results have provided unprecedented profiles of the cytokines (TNF, IL-1, IL-6, GCSF, and MIF), chemokines (MIP-1α, MIP-1β, IL-8, and Gro-1), certain transcription factors, and the matrix metalloproteinases (GelA, Strom-1, Col-1, Col-3, HME) in the macrophage cell line MM6 and in the SW1353 chondrosarcoma cells.

Earlier reports of cytokine production in the diseased state had established a model in which TNF is a major participant in RA. Its expression reportedly preceded that of the other cytokines and effector molecules (4). Our results strongly support these results as demonstrated in the time course of the MM6 cells where TNF induction preceded that of IL-1α and IL-β followed by IL-6 and GCSF. These expression profiles demonstrate the utility of the microarrays in determining the hierarachy of signaling events.

In the SW1353 chondrosarcoma cells, all the known MMPs and TIMPs were examined simultaneously. HME expression was discovered, which previously had been observed in only the stromal cells and alveolar macrophages of smoker's lungs and in placental tissue. Its presence in cells of the RA tissue is meaningful because its activity can cause significant destruction of elastin and basement membrane components (16, 17). Expression profiles of synovial fibroblasts and articular chondrocytes were remarkably similar and not too different from the SW1353 cells, indicating that the fibroblast and the chondrocyte can play equally aggressive roles in joint erosion. Prominent genes expressed were

FIG. 5.   Expression profiles of RA tissue (*A*) and IBD tissue (*B*). mRNA from RA tissue samples obtained from the same individual was isolated directly after excision (RA 21.5A) or maintained in culture without serum for 2 hr (RA 21.5B) or for 6 hr (RA 21.5C). Profiles from tissue samples of two other individuals (data not shown) were remarkably similar to the ones shown here. IBD-A and IBD-CI are from mRNA samples prepared directly after surgery from two separate individuals. For the IBD-CII probe, the tissue sample was cultured in medium without serum for 2 hr before mRNA preparation.

the MMPs, but chemokines and cytokines were also produced by these cells. The effect of the anabolic growth factor TGF-β was profoundly evident in demonstrating the down regulation of these catabolic activities.

RA tissue samples undeniably reflected profiles similar to the cell types examined. Active genes observed were IL-3, IL-6, ICE, the MMPs including HME and TIMPs, chemokines IL-8, Groα, MIP, MIF, and RANTES, and the adhesion molecule VCAM. Of the growth factors, fibroblast growth factor β was observed most frequently. In comparison, the expression patterns in the other inflammatory state (i.e., IBD) were not as marked as in the RA samples, at least as obtained from the tissue samples selected for this study.

As an alternative approach, the 1046 cDNA microarray of randomly selected genes from a lymphocyte library was used to identify genes expressed in RA tissue (3). Many genes on this array hybridized with probes made from both RA and IBD tissue samples. The results are not surprising because inflammatory tissue is abundantly supplied with cell types infiltrating from the circulating blood, made apparent also by the high levels of chemokine expression in RA tissue. Because of the magnitude of the effort required to identify all the hybridized genes, we have for this report chosen to describe only three differentially expressed genes mainly to verify this method of analysis.

Of the large number of genes observed here, a fair number were already known as active participants in inflammatory disease. These are TNF, IL-1, IL-6, IL-8, GCSF, RANTES, and VCAM. The novel participants not previously reported are HME, IL-3, ICE, and Groα. With our discovery of HME expression in RA, this gene becomes a target for drug intervention. ICE is a cysteine protease well known for its IL-1β processing activity (18), and recognized for its role in apoptotic cell death (19). Its expression in RA tissue is intriguing. IL-3 is recognized for its growth-promoting activity in hematopoietic cell lineages, is a product of activated T cells (20), and its expression in synoviocytes and chondrocytes of RA tissue is a novel observation.

Like IL-8, Groα, is a C-X-C subgroup chemokine and is a potent neutrophil and basophil chemoattractant. It downregulates the expression of types I and III interstitial collagens (21, 22) and is seen here produced by the MM6 cells, in primary synoviocytes, and in RA tissue. With the presence of RANTES, MCP, and MIP-1β, the C-C chemokines (23) migration and infiltration of monocytes, particularly T cells, into the tissue is also enhanced (5) and aid in the trafficking and recruitment of leukocytes into the RA tissue. Their activation, phagocytosis, degranulation, and respiratory bursts could be responsible for the induction of MnSOD in RA. MnSOD is also induced by TNF and IL-1 and serves a protective function against oxidative damage. The induction of the ferritin light chain encoding gene in this tissue may be for reasons similar to those for MnSOD. Ferritin is the major intracellular iron storage protein and it is responsive to intracellular oxidative stress and reactive oxygen intermediates generated during inflammation (24, 25). The active expression of TIMP-1 in RA tissue, as detected by the 1000-element array, is no surprise because our results have repeatedly shown TIMP-1 to be expressed in the constitutive and induced states of RA cells and tissues.

The suitability of the cDNA microarray technology for profiling diseases and for identifying disease related genes is well documented here. This technology could provide new targets for drug development and disease therapies, and in doing so allow for improved treatment of chronic diseases that are challenging because of their complexity.

1. Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995) *Science* **270**, 467–470.
2. Shalon, D., Smith, S. & Brown, P. O. (1996) *Genome Res.* **6**, 639–645.
3. Schena, M., Shalon, D., Heller, R., Chai, A., Brown, P. O. & Davis, R. W. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 10614–10619.
4. Feldmann, M., Brennan F. M. & Maini, R. N. (1996) *Rheumatoid Arthritis Cell* **85**, 307–310.
5. Schall, T. J. (1994) in *The Cytokine Handbook*, ed. Thomson, A. W. (Academic, New York), 2nd Ed., pp. 410–460.
6. Lotz, M. F., Blanco, J., Von Kempis, J., Dudler, J., Maier, R., Villiger P. M. & Geng, Y. (1995) *J. Rheumatol.* **22**, Supplement 43, 104–108.
7. Birkedal-Hansen, H., Moore, W. G. I., Bodden, M. K., Windsor, L. J., Birkedal-Hansen, B., DeCarlo, A. &. Engler, J. A. (1993) *Crit. Rev. Oral Biol. Med.* **4**, 197–250.
8. Zeigler-Heitbrock, H. W. L., Thiel, E., Futterer, A., Volker, H., Wirtz, A. & Reithmuller, G. (1988) *Int. J. Cancer* **41**, 456–461.
9. Borden, P., Solymar, D., Sucharczuk, A., Lindman, B., Cannon, P. & Heller, R. A. (1996) *J. Biol. Chem.* **271**, 23577–23581.
10. Gadher, S. J. & Woolley, D. E. (1987) *Rheumatol. Int.* **7**, 13–22.
11. Harris, E. D., Jr. (1990) *New Engl. J. Med.* **322**, 1277–1289.
12. Firestein, G. S. (1996) in *Textbook of Rheumatology*, eds. Kelly, W. N., Harris, E. D., Ruddy, S. & Sledge, C. B. (Saunders, Philadelphia), 5th Ed. pp. 5001–5047.
13. Alvaro-Garcia, J. M., Zvaiflfer, Nathan J., Brown, C. B., Kaushansky, K. & Firestein, Gary S. (1991) *J. Immunol.* **146**, 3365–3371.
14. Firestein, G. S., Alvaro-Grarcia, J. M. & Maki, R. (1990) *J. Immunol.* **144**, 3347–3352.
15. Pradines-Figueres, A. & Raetz, C. R. H. (1992) *J. Biol. Chem.* **267**, 23261–23268.
16. Shapiro, S. D., Kobayashi, D. L. & Ley, T. J. (1993) *J. Biol. Chem.* **208**, 23824–23829.
17. Shipley, M. J., Wesselschmidt, R. L., Kobayashi, D. K., Ley, T. J. & Shapiro, S. D. (1996) *Proc. Natl. Acad. Sci. USA* **93**, 3042–3946.
18. Cerreti, D. P., Kozlosky, C. J., Mosley, B., Nelson, N., Van Ness, K., Greenstreet, T. A., March, C. J., Kronheim, S. R., Druck, T., Cannizaro, L. A., Huebner, K. & Black, R. A. (1992) *Science* **256**, 97–100.
19. Miura, M., Zhu, H., Rotello, R., Hartweig, E. A. & Yuan, J. (1993) *Cell* **75**, 653–660.
20. Arai, K., Lee, F., Miyajima, A., Shoichiro, M., Arai, N. & Takashi, Y. (1990) *Annu. Rev. Biochem.* **59**, 783–836.
21. Geiser, T., Dewald, B., Ehrengruber, M. U., Lewis, I. C. & Baggiolini, M. (1993) *J. Biol. Chem.* **268**, 15419–15424.
22. Unemori, E. N., Amento, E. P., Bauer, E. A. & Horuk, R. (1993) *J. Biol. Chem.* **268**, 1338–1342.
23. Robinson, E., Keystone, E. C., Schall, T. J., Gillet, N. & Fish, E. N. (1995) *Clin. Exp. Immunol.* **101**, 398–407.
24. Roeser, H. (1980) in *Iron Metabolism in Biochemistry and Medicine*, eds. Jacobs, A. & Worwood, M. (Academic, New York), Vol. 2, pp. 605–640.
25. Kwak, E. L., Larochelle, D. A., Beaumont, C., Torti, S. V. & Torti, F. M. (1995) *J. Biol. Chem.* **270**, 15285–15293.

PCT

INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(51) International Patent Classification 6 : **C12Q 1/68, C07H 21/04**	**A1**	(11) International Publication Number: **WO 97/13877**
		(43) International Publication Date: 17 April 1997 (17.04.97)

(21) International Application Number: PCT/US96/16342

(22) International Filing Date: 11 October 1996 (11.10.96)

(30) Priority Data:
PCT/US95/12791 12 October 1995 (12.10.95) WO
(34) Countries for which the regional or
international application was filed: US et al.
PCT/US96/09513 6 June 1996 (06.06.96) WO
(34) Countries for which the regional or
international application was filed: US et al.

(60) Parent Application or Grant
(63) Related by Continuation
US Not furnished (CIP)
Filed on Not furnished

(71) Applicant (for all designated States except US): LYNX THER-APEUTICS, INC. [US/US]; 3832 Bay Center Place, Hayward, CA 94545 (US).

(72) Inventor; and
(75) Inventor/Applicant (for US only): MARTIN, David, W. [US/US]; Lynx Therapeutics, Inc., 3832 Bay Center Place, Hayward, CA 94545 (US).

(74) Agent: POWERS, Vincent, M.; Dehlinger & Associates, Post Office Box 60850, Palo Alto, CA 94306-0850 (US).

(81) Designated States: AU, CA, CZ, EE, FI, HU, JP, KR, LT, LV, NO, NZ, PL, RU, SG, US, European patent (AT, BE, CH, DE, DK, ES, FI, FR, GB, GR, IE, IT, LU, MC, NL, PT, SE).

Published
With international search report.
Before the expiration of the time limit for amending the claims and to be republished in the event of the receipt of amendments.

(54) Title: MEASUREMENT OF GENE EXPRESSION PROFILES IN TOXICITY DETERMINATION

(57) Abstract

A method is provided for assessing the toxicity of a compound in a test organism by measuring gene expression profiles of selected tissues. Gene expression profiles are measured by massively parallel signature sequencing of cDNA libraries constructed from mRNA extracted from the selected tissues. Gene expression profiles provide extensive information on the effects of administering a compound to a test organism in both acute toxicity tests and in prolonged and chronic toxicity tests.

MEASUREMENT OF GENE EXPRESSION PROFILES
IN TOXICITY DETERMINATION

5   Field of the Invention

The invention relates generally to methods for detecting and monitoring phenotypic changes in in vitro and in vivo systems for assessing and/or determining the toxicity of chemical compounds, and more particularly, the invention relates to a method for detecting and monitoring changes in gene expression patterns in in vitro

10   and in vivo systems for determining the toxicity of drug candidates.

BACKGROUND

The ability to rapidly and conveniently assess the toxicity of new compounds is extremely important. Thousands of new compounds are synthesized every year,

15   and many are introduced to the environment through the development of new commercial products and processes, often with little knowledge of their short term and long term health effects. In the development of new drugs, the cost of assessing the safety and efficacy of candidate compounds is becoming astronomical: It is estimated that the pharmaceutical industry spends an average of about 300 million

20   dollars to bring a new pharmaceutical compound to market, e.g. Biotechnology, 13: 226-228 (1995). A large fraction of these costs are due to the failure of candidate compounds in the later stages of the developmental process. That is, as the assessment of a candidate drug progresses from the identification of a compound as a drug candidate--for example, through relatively inexpensive binding assays or in vitro

25   screening assays, to pharmacokinetic studies, to toxicity studies, to efficacy studies in model systems, to preliminary clinical studies, and so on, the costs of the associated tests and analyses increases tremendously. Consequently, it may cost several tens of millions of dollars to determine that a once promising candidate compound possesses a side effect or cross reactivity that renders it commercially infeasible to develop

30   further. A great challenge of pharmaceutical development is to remove from further consideration as early as possible those compounds that are likely to fail in the later stages of drug testing.

Drug development programs are clearly structured with this objective in mind; however, rapidly escalating costs have created a need to develop even more stringent

35   and less expensive screens in the early stages to identify false leads as soon as possible. Toxicity assessment is an area where such improvements may be made, for both drug development and for assessing the environmental, health, and safety effects of new compounds in general.

Typically the toxicity of a compound is determined by administering the compound to one or more species of test animal under controlled conditions and by monitoring the effects on a wide range of parameters. The parameters include such things as blood chemistry, weight gain or loss, a variety of behavioral patterns, muscle

5   tone, body temperature, respiration rate, lethality, and the like, which collectively provide a measure of the state of health of the test animal. The degree of deviation of such parameters from their normal ranges gives a measure of the toxicity of a compound. Such tests may be designed to assess the acute, prolonged, or chronic toxicity of a compound. In general, acute tests involve administration of the test

10   chemical on one occasion. The period of observation of the test animals may be as short as a few hours, although it is usually at least 24 hours and in some cases it may be as long as a week or more. In general, prolonged tests involve administration of the test chemical on multiple occasions. The test chemical may be administered one or more times each day, irregularly as when it is incorporated in the diet, at specific

15   times such as during pregnancy, or in some cases regularly but only at weekly intervals. Also, in the prolonged test the experiment is usually conducted for not less than 90 days in the rat or mouse or a year in the dog. In contrast to the acute and prolonged types of test, the chronic toxicity tests are those in which the test chemical .is administered for a substantial portion of the lifetime of the test animal. In the case

20   of the mouse or rat, this is a period of 2 to 3 years. In the case of the dog, it is for 5 to 7 years.

Significant costs are incurred in establishing and maintaining large cohorts of test animals for such assays, especially the larger animals in chronic toxicity assays. Moreover, because of species specific effects, passing such toxicity tests does not

25   ensure that a compound is free of toxic effects when used in humans. Such tests do, however, provide a standardized set of information for judging the safety of new compounds, and they provide a database for giving preliminary assessments of related compounds. An important area for improving toxicity determination would be the identification of new observables which are predictive of the outcome of the

30   expensive and tedious animal assays.

In other medical fields, there has been significant interest in applying recent advances in biotechnology, particularly in DNA sequencing, to the identification and study of differentially expressed genes in healthy and diseased organisms, e.g. Adams et al, Science, 252: 1651-1656 (1991); Matsubara et al, Gene, 135: 265-274 (1993);

35   Rosenberg et al, International patent application, PCT/US95/01863. The objectives of such applications include increasing our knowledge of disease processes, identifying genes that play important roles in the disease process, and providing diagnostic and therapeutic approaches that exploit the expressed genes or their

products. While such approaches are attractive, those based on exhaustive, or even sampled, sequencing of expressed genes are still beset by the enormous effort required: It is estimated that 30-35 thousand different genes are expressed in a typical mammalian tissue in any given state, e.g. Ausubel et al, Editors, Current Protocols,

5   5.8.1-5.8.4 (John Wiley & Sons, New York, 1992). Determining the sequences of even a small sample of that number of gene products is a major enterprise, requiring industrial-scale resources. Thus, the routine application of massive sequencing of expressed genes is still beyond current commercial technology.

The availability of new assays for assessing the toxicity of compounds, such

10   as candidate drugs, that would provide more comprehensive and precise information about the state of health of a test animal would be highly desirable. Such additional assays would preferably be less expensive, more rapid, and more convenient than current testing procedures, and would at the same time provide enough information to make early judgments regarding the safety of new compounds.

15

Summary of the Invention

An object of the invention is to provide a new approach to toxicity assessment based on an examination of gene expression patterns, or profiles, in in vitro or in vivo test systems.

20   Another object of the invention is to provide a database on which to base decisions concerning the toxicological properties of chemicals, particularly drug candidates.

A further object of the invention is to provide a method for analyzing gene expression patterns in selected tissues of test animals.

25   A still further object of the invention is to provide a system for identifying genes which are differentially expressed in response to exposure to a test compound.

Another object of the invention is to provide a rapid and reliable method for correlating gene expression with short term and long term toxicity in test animals.

Another object of the invention is to identify genes whose expression is

30   predictive of deleterious toxicity.

The invention achieves these and other objects by providing a method for massively parallel signature sequencing of genes expressed in one or more selected tissues of an organism exposed to a test compound. An important feature of the invention is the application of novel DNA sorting and sequencing methodologies that

35   permit the formation of gene expression profiles for selected tissues by determining the sequence of portions of many thousands of different polynucleotides in parallel. Such profiles may be compared with those from tissues of control organisms at single or multiple time points to identify expression patterns predictive of toxicity.

The sorting methodology of the invention makes use of oligonucleotide tags that are members of a minimally cross-hybridizing set of oligonucleotides. The sequences of oligonucleotides of such a set differ from the sequences of every other member of the same set by at least two nucleotides. Thus, each member of such a set

5   cannot form a duplex (or triplex) with the complement of any other member with less than two mismatches. Complements of oligonucleotide tags of the invention, referred to herein as "tag complements," may comprise natural nucleotides or non-natural nucleotide analogs. Preferably, tag complements are attached to solid phase supports. Such oligonucleotide tags when used with their corresponding tag complements

10   provide a means of enhancing specificity of hybridization for sorting polynucleotides, such as cDNAs.

     The polynucleotides to be sorted each have an oligonucleotide tag attached, such that different polynucleotides have different tags. As explained more fully below, this condition is achieved by employing a repertoire of tags substantially

15   greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed under conditions which permit specific hybridization of the oligonucleotide tags with -their respective complements, identical polynucleotides sort onto particular beads or

20   regions. The sorted populations of polynucleotides can then be sequenced on the solid phase support by a "single-base" or "base-by-base" sequencing methodology, as described more fully below.

     In one aspect, the method of the invention comprises the following steps: (a) administering the compound to a test organism; (b) extracting a population of mRNA

25   molecules from each of one or more tissues of the test organism; (c) forming a separate population of cDNA molecules from each population of mRNA molecules extracted from the one or more tissues such that each cDNA molecule of the separate populations has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set; (d) separately sampling each

30   population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached; (e) sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in

35   spatially discrete regions on one or more solid phase supports; (f) determining the nucleotide sequence of a portion of each of the sorted cDNA molecules of each separate population to form a frequency distribution of expressed genes for each of

the one or more tissues; and (g) correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

An important aspect of the invention is the identification of genes whose expression is predictive of the toxicity of a compound. Once such genes are

5   identified, they may be employed in conventional assays, such as reverse transcriptase polymerase chain reaction (RT-PCR) assays for gene expression.

Brief Description of the Drawings

Figure 1 is a flow chart representation of an algorithm for generating

10   minimally cross-hybridizing sets of oligonucleotides.

Figure 2 diagrammatically illustrates an apparatus for carrying out polynucleotide sequencing in accordance with the invention.

Definitions

15   "Complement" or "tag complement" as used herein in reference to oligonucleotide tags refers to an oligonucleotide to which a oligonucleotide tag specifically hybridizes to form a perfectly matched duplex or triplex. In embodiments where specific hybridization results in a triplex, the oligonucleotide tag may be _selected to be either double stranded or single stranded. Thus, where triplexes are

20   formed, the term "complement" is meant to encompass either a double stranded complement of a single stranded oligonucleotide tag or a single stranded complement of a double stranded oligonucleotide tag.

The term "oligonucleotide" as used herein includes linear oligomers of natural or modified monomers or linkages, including deoxyribonucleosides, ribonucleosides,

25   anomeric forms thereof, peptide nucleic acids (PNAs), and the like, capable of specifically binding to a target polynucleotide by way of a regular pattern of monomer-to-monomer interactions, such as Watson-Crick type of base pairing, base stacking, Hoogsteen or reverse Hoogsteen types of base pairing, or the like. Usually monomers are linked by phosphodiester bonds or analogs thereof to form

30   oligonucleotides ranging in size from a few monomeric units, e.g. 3-4, to several tens of monomeric units. Whenever an oligonucleotide is represented by a sequence of letters, such as "ATGCCTG," it will be understood that the nucleotides are in 5'→3' order from left to right and that "A" denotes deoxyadenosine, "C" denotes deoxycytidine, "G" denotes deoxyguanosine, and "T" denotes thymidine, unless

35   otherwise noted. Analogs of phosphodiester linkages include phosphorothioate, phosphorodithioate, phosphoranilidate, phosphoramidate, and the like. Usually oligonucleotides of the invention comprise the four natural nucleotides; however, they may also comprise non-natural nucleotide analogs. It is clear to those skilled in the

art when oligonucleotides having natural or non-natural nucleotides may be employed, e.g. where processing by enzymes is called for, usually oligonucleotides consisting of natural nucleotides are required.

"Perfectly matched" in reference to a duplex means that the poly- or
5   oligonucleotide strands making up the duplex form a double stranded structure with one other such that every nucleotide in each strand undergoes Watson-Crick basepairing with a nucleotide in the other strand. The term also comprehends the pairing of nucleoside analogs, such as deoxyinosine, nucleosides with 2-aminopurine bases, and the like, that may be employed. In reference to a triplex, the term means
10   that the triplex consists of a perfectly matched duplex and a third strand in which every nucleotide undergoes Hoogsteen or reverse Hoogsteen association with a basepair of the perfectly matched duplex. Conversely, a "mismatch" in a duplex between a tag and an oligonucleotide means that a pair or triplet of nucleotides in the duplex or triplex fails to undergo Watson-Crick and/or Hoogsteen and/or reverse
15   Hoogsteen bonding.

As used herein, "nucleoside" includes the natural nucleosides, including 2'-deoxy and 2'-hydroxyl forms, e.g. as described in Kornberg and Baker, DNA Replication, 2nd Ed. (Freeman, San Francisco, 1992). "Analogs" in reference to nucleosides includes synthetic nucleosides having modified base moieties and/or
20   modified sugar moieties, e.g. described by Scheit, Nucleotide Analogs (John Wiley, New York, 1980); Uhlman and Peyman, Chemical Reviews, 90: 543-584 (1990). or the like, with the only proviso that they are capable of specific hybridization. Such analogs include synthetic nucleosides designed to enhance binding properties, reduce complexity, increase specificity, and the like.

25   As used herein "sequence determination" or "determining a nucleotide sequence" in reference to polynucleotides includes determination of partial as well as full sequence information of the polynucleotide. That is, the term includes sequence comparisons, fingerprinting, and like levels of information about a target polynucleotide, as well as the express identification and ordering of nucleosides,
30   usually each nucleoside, in a target polynucleotide. The term also includes the determination of the identification, ordering, and locations of one, two, or three of the four types of nucleotides within a target polynucleotide. For example, in some embodiments sequence determination may be effected by identifying the ordering and locations of a single type of nucleotide. e.g. cytosines, within the target polynucleotide
35   "CATCGC ..." so that its sequence is represented as a binary code, e.g. "100101 ... " for "C-(not C)-(not C)-C-(not C)-C ... " and the like.

As used herein, the term "complexity" in reference to a population of polynucleotides means the number of different species of molecule present in the population.

5   As used herein, the terms "gene expression profile," and "gene expression pattern" which is used equivalently, means a frequency distribution of sequences of portions of cDNA molecules sampled from a population of tag-cDNA conjugates. Generally, the portions of sequence are sufficiently long to uniquely identify the cDNA from which the portion arose. Preferably, the total number of sequences determined is at least 1000; more preferably, the total number of sequences

10   determined in a gene expression profile is at least ten thousand.

As used herein, "test organism" means any in vitro or in vivo system which provides measureable responses to exposure to test compounds. Typically, test organisms may be mammalian cell cultures, particularly of specific tissues, such as hepatocytes, neurons, kidney cells, colony forming cells, or the like, or test organisms

15   may be whole animals, such as rats, mice, hamsters, guinea pigs, dogs, cats, rabbits, pigs, monkeys, and the like.

Detailed Description of the Invention

The invention provides a method for determining the toxicity of a compound

20   by analyzing changes in the gene expression profiles in selected tissues of test organisms exposed to the compound. The invention also provides a method of identifying toxicity markers consisting of individual genes or a group of genes that is expressed acutely and which is correlated with prolonged or chronic toxicity, or suggests that the compound will have an undesirable cross reactivity. Gene

25   expression profiles are generated by sequencing portions of cDNA molecules construction from mRNA extracted from tissues of test organisms exposed to the compound being tested. As used herein, the term "tissue" is employed with its usual medical or biological meaning, except that in reference to an in vitro test system, such as a cell culture, it simply means a sample from the culture. Gene expression profiles

30   derived from test organisms are compared to gene expression profiles derived from control organisms to determine the genes which are differentially expressed in the test organism because of exposure to the compound being tested. In both cases, the sequence information of the gene expression profiles is obtained by massively parallel signature sequencing of cDNAs, which is implemented in steps (c) through (f) of the

35   above method.

Toxicity Assessment

Procedures for designing and conducting toxicity tests in in vitro and in vivo systems is well known, and is described in many texts on the subject, such as Loomis

et al. Loomis's Esstentials of Toxicology, 4th Ed. (Academic Press, New York, 1996); Echobichon, The Basics of Toxicity Testing (CRC Press, Boca Raton, 1992); Frazier, editor, In Vitro Toxicity Testing (Marcel Dekker, New York, 1992); and the like.

In toxicity testing, two groups of test organisms are usually employed: one group serves as a control and the other group receives the test compound in a single dose (for acute toxicity tests) or a regimen of doses (for prolonged or chronic toxicity tests). Since in most cases, the extraction of tissue as called for in the method of the invention requires sacrificing the test animal, both the control group and the group receiving compound must be large enough to permit removal of animals for sampling tissues, if it is desired to observe the dynamics of gene expression through the duration of an experiment.

In setting up a toxicity study, extensive guidance is provided in the literature for selecting the appropriate test organism for the compound being tested, route of administration, dose ranges, and the like. Water or physiological saline (0.9% NaCl in water) is the solute of choice for the test compound since these solvents permit administration by a variety of routes. When this is not possible because of solubility limitations, it is necessary to resort to the use of vegetable oils such as corn oil or even organic solvents, of which propylene glycol is commonly used. Whenever -possible the use of suspension of emulsion should be avoided except for oral administration. Regardless of the route of administration, the volume required to administer a given dose is limited by the size of the animal that is used. It is desirable to keep the volume of each dose uniform within and between groups of animals. When rates or mice are used the volume administered by the oral route should not exceed 0.005 ml per gram of animal. Even when aqueous or physiological saline solutions are used for parenteral injection the volumes that are tolerated are limited, although such solutions are ordinarily thought of as being innocuous. The intravenous LD_{50} of distilled water in the mouse is approximately 0.044 ml per gram and that of isotonic saline is 0.068 ml per gram of mouse.

When a compound is to be administered by inhalation, special techniques for generating test atmospheres are necessary. Dose estimation becomes very complicated. The methods usually involve aerosolization or nebulization of fluids containing the compound. If the agent to be tested is a fluid that has an appreciable vapor pressure, it may be administered by passing air through the solution under controlled temperature conditions. Under these conditions, dose is estimated from the volume of air inhaled per unit time, the temperature of the solution, and the vapor pressure of the agent involved. Gases are metered from reservoirs. When particles of a solution are to be administered. unless the particle size is less than about 2 μm the particles will not reach the terminal alveolar sacs in the lungs. A variety of

apparatuses and chambers are available to perform studies for detecting effects of irritant or other toxic endpoints when they are administered by inhalation. The preferred method of administering an agent to animals is via the oral route, either by intubation or by incorporating the agent in the feed.

5   Preferably, in designing a toxicity assessment, two or more species should be employed that handle the test compound as similarly to man as possible in terms of metabolism, absorption, excretion, tissue storage, and the like. Preferably, multiple doses or regimens at different concentrations should be employed to establish a dose-response relationship with respect to toxic effects. And preferably, the route of

10   administration to the test animal should be the same as, or as similar as possible to, the route of administration of the compound to man. Effects obtained by one route of administration to test animals are not a priori applicable to effects by another route of administration to man. For example, food additives for man should be tested by admixture of the material in the diet of the test animals.

15   Acute toxicity tests consist of administering a compound to test organisms on one occasion. The purpose of such test is to determine the symptomotology consequent to administration of the compound and to determine the degree of lethality of the compound. The initial procedure is to perform a series of range-finding doses .of the compound in a single species. This necessitates selection of a route of

20   administration, preparation of the compound in a form suitable for administration by the selected route, and selection of an appropriate species. Preferably, initial acute toxicity studies are performed on either rats or mice because of their low cost, their availability, and the availability of abundant toxicologic reference data on these species. Prolonged toxicity tests consist of administering a compound to test

25   organisms repeatedly, usually on a daily basis, over a period of 3 to 4 months. Two practical factors are encountered that place constraints on the design of such tests: First, the available routes of administration are limited because the route selected must be suitable for repeated administration without inducing harmful effects. And second, blood, urine, and perhaps other samples, should be taken repeatedly without

30   inducing significant harm to the test animals. Preferably, in the method of the · invention the gene expression profiles are obtained in conjunction with the measurement of the traditional toxicologic parameters, such as listed in the table below:

35

Hematology	Blood Chemistry	Urine Analyses
erythrocyte count	sodium	pH
total leukocyte count	potassium	specific gravity
differential leukocyte count	chloride	total protein
hematocrit	calcium	sediment
hemoglobin	carbon dioxide	glucose
	serum glutamine-pyruvate transaminase	ketones
	serum glutamin-oxalacetic transaminase	bilirubin
	serum protein electrophoresis	
	blood sugar	
	blood urea nitrogen	
	total serum protein	
	serum albumin	
	total serum bilirubin	

5   Oligonucleotide Tags and Tag Complements

Oligonucleotide tags are members of a minimally cross-hybridizing set of
oligonucleotides. The sequences of oligonucleotides of such a set differ from the
sequences of every other member of the same set by at least two nucleotides. Thus,
each member of such a set cannot form a duplex (or triplex) with the complement of
10   any other member with less than two mismatches. Complements of oligonucleotide
tags, referred to herein as "tag complements," may comprise natural nucleotides or
non-natural nucleotide analogs. Preferably, tag complements are attached to solid
phase supports. Such oligonucleotide tags when used with their corresponding tag
complements provide a means of enhancing specificity of hybridization for sorting,
15   tracking, or labeling molecules, especially polynucleotides.

Minimally cross-hybridizing sets of oligonucleotide tags and tag complements
may be synthesized either combinatorially or individually depending on the size of the
set desired and the degree to which cross-hybridization is sought to be minimized (or
stated another way, the degree to which specificity is sought to be enhanced). For
20   example, a minimally cross-hybridizing set may consist of a set of individually
synthesized 10-mer sequences that differ from each other by at least 4 nucleotides.
such set having a maximum size of 332 (when composed of 3 kinds of nucleotides
and counted using a computer program such as disclosed in Appendix Ic).
Alternatively, a minimally cross-hybridizing set of oligonucleotide tags may also be

assembled combinatorially from subunits which themselves are selected from a
minimally cross-hybridizing set. For example, a set of minimally cross-hybridizing
12-mers differing from one another by at least three nucleotides may be synthesized
by assembling 3 subunits selected from a set of minimally cross-hybridizing 4-mers
5   that each differ from one another by three nucleotides. Such an embodiment gives a
maximally sized set of 9^3, or 729, 12-mers. The number 9 is number of
oligonucleotides listed by the computer program of Appendix Ia, which assumes, as
with the 10-mers, that only 3 of the 4 different types of nucleotides are used. The set
is described as "maximal" because the computer programs of Appendices Ia-c provide
10   the largest set for a given input (e.g. length, composition, difference in number of
nucleotides between members). Additional minimally cross-hybridizing sets may be
formed from subsets of such calculated sets.

      Oligonucleotide tags may be single stranded and be designed for specific
hybridization to single stranded tag complements by duplex formation or for specific
15   hybridization to double stranded tag complements by triplex formation.
Oligonucleotide tags may also be double stranded and be designed for specific
hybridization to single stranded tag complements by triplex formation.

      When synthesized combinatorially, an oligonucleotide tag preferably consists
-of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9
20   nucleotides in length wherein each subunit is selected from the same minimally cross-
hybridizing set. In such embodiments, the number of oligonucleotide tags available
depends on the number of subunits per tag and on the length of the subunits. The
number is generally much less than the number of all possible sequences the length of
the tag, which for a tag n nucleotides long would be 4^n.

25   Complements of oligonucleotide tags attached to a solid phase support are
used to sort polynucleotides from a mixture of polynucleotides each containing a tag.
Complements of the oligonucleotide tags are synthesized on the surface of a solid
phase support, such as a microscopic bead or a specific location on an array of
synthesis locations on a single support, such that populations of identical sequences
30   are produced in specific regions. That is, the surface of each support, in the case of a
bead, or of each region, in the case of an array, is derivatized by only one type of
complement which has a particular sequence. The population of such beads or regions
contains a repertoire of complements with distinct sequences. As used herein in
reference to oligonucleotide tags and tag complements, the term "repertoire" means
35   the set of minimally cross-hybridizing set of oligonucleotides that make up the tags in
a particular embodiment or the corresponding set of tag complements.

      The polynucleotides to be sorted each have an oligonucleotide tag attached,
such that different polynucleotides have different tags. As explained more fully ·

below, this condition is achieved by employing a repertoire of tags substantially greater than the population of polynucleotides and by taking a sufficiently small sample of tagged polynucleotides from the full ensemble of tagged polynucleotides. After such sampling, when the populations of supports and polynucleotides are mixed

5   under conditions which permit specific hybridization of the oligonucleotide tags with their respective complements, identical polynucleotides sort onto particular beads or regions.

The nucleotide sequences of oligonucleotides of a minimally cross-hybridizing set are conveniently enumerated by simple computer programs, such as those

10   exemplified by programs whose source codes are listed in Appendices Ia and Ib. Program minhx of Appendix Ia computes all minimally cross-hybridizing sets having 4-mer subunits composed of three kinds of nucleotides. Program tagN of Appendix Ib enumerates longer oligonucleotides of a minimally cross-hybridizing set. Similar algorithms and computer programs are readily written for listing oligonucleotides of

15   minimally cross-hybridizing sets for any embodiment of the invention. Table I below provides guidance as to the size of sets of minimally cross-hybridizing oligonucleotides for the indicated lengths and number of nucleotide differences. The above computer programs were used to generate the numbers.

20   Table I

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross-Hybridizing Set	Maximal Size of Minimally Cross-Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	3	9	6561	5.90×10^4
6	3	27	5.3×10^5	1.43×10^7
7	4	27	5.3×10^5	1.43×10^7
7	5	8	4096	3.28×10^4
8	3	190	1.30×10^9	2.48×10^{11}
8	4	62	1.48×10^7	9.16×10^8
8	5	18	1.05×10^5	1.89×10^6
9	5	39	2.31×10^6	9.02×10^7
10	5	332	1.21×10^{10}	
10	6	28	6.15×10^5	1.72×10^7
11	5	187		
18	6	≈25000		

18   12   24

For some embodiments of the invention, where extremely large repertoires of tags are not required, oligonucleotide tags of a minimally cross-hybridizing set may be separately synthesized. Sets containing several hundred to several thousands, or even several tens of thousands, of oligonucleotides may be synthesized directly by a variety of parallel synthesis approaches, e.g. as disclosed in Frank et al, U.S. patent 4,689,405; Frank et al, Nucleic Acids Research, 11: 4365-4377 (1983); Matson et al, Anal. Biochem., 224: 110-116 (1995); Fodor et al, International application PCT/US93/04145; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern et al, J. Biotechnology, 35: 217-227 (1994), Brennan, International application PCT/US94/05896; Lashkari et al, Proc. Natl. Acad. Sci., 92: 7912-7915 (1995); or the like.

Preferably, oligonucleotide tags of the invention are synthesized combinatorially out of subunits between three and six nucleotides in length and selected from the same minimally cross-hybridizing set. For oligonucletides in this range, the members of such sets may be enumerated by computer programs based on the algorithm of Fig. 1.

The algorithm of Fig. 1 is implemented by first defining the characteristics of the subunits of the minimally cross-hybridizing set, i.e. length, number of base differences between members, and composition, e.g. do they consist of two, three, or four kinds of bases. A table M_n, n=1, is generated (100) that consists of all possible sequences of a given length and composition. An initial subunit S_1 is selected and compared (120) with successive subunits S_i for i=n+1 to the end of the table. Whenever a successive subunit has the required number of mismatches to be a member of the minimally cross-hybridizing set, it is saved in a new table M_{n+1} (125). that also contains subunits previously selected in prior passes through step 120. For example, in the first set of comparisons, M_2 will contain S_1; in the second set of comparisons, M_3 will contain S_1 and S_2; in the third set of comparisons, M_4 will contain S_1, S_2, and S_3; and so on. Similarly, comparisons in table M_j will be between S_j and all successive subunits in M_j. Note that each successive table M_{n+1} is smaller than its predecessors as subunits are eliminated in successive passes through step 130. After every subunit of table M_n has been compared (140) the old table is replaced by the new table M_{n+1}, and the next round of comparisons are begun. The process stops (160) when a table M_n is reached that contains no successive subunits to compare to the selected subunit S_i, i.e. $M_n = M_{n+1}$.

Preferably, minimally cross-hybridizing sets comprise subunits that make approximately equivalent contributions to duplex stability as every other subunit in

the set. In this way, the stability of perfectly matched duplexes between every subunit and its complement is approximately equal. Guidance for selecting such sets is provided by published techniques for selecting optimal PCR primers and calculating duplex stabilities, e.g. Rychlik et al, Nucleic Acids Research, 17: 8543-8551 (1989)

5   and 18: 6409-6412 (1990); Breslauer et al, Proc. Natl. Acad. Sci., 83: 3746-3750 (1986); Wetmur, Crit. Rev. Biochem. Mol. Biol., 26: 227-259 (1991);and the like. For shorter tags, e.g. about 30 nucleotides or less, the algorithm described by Rychlik and Wetmur is preferred, and for longer tags, e.g. about 30-35 nucleotides or greater, an algorithm disclosed by Suggs et al, pages 683-693 in Brown, editor, ICN-UCLA

10   Symp. Dev. Biol., Vol. 23 (Academic Press, New York, 1981) may be conveniently employed. Clearly, the are many approaches available to one skilled in the art for designing sets of minimally cross-hybridizing subunits within the scope of the invention. For example, to minimize the affects of different base-stacking energies of terminal nucleotides when subunits are assembled, subunits may be provided that

15   have the same terminal nucleotides. In this way, when subunits are linked, the sum of the base-stacking energies of all the adjoining terminal nucleotides will be the same, thereby reducing or eliminating variability in tag melting temperatures.

A "word" of terminal nucleotides, shown in italic below, may also be added to each end of a tag so that a perfect match is always formed between it and a similar

20   terminal "word" on any other tag complement. Such an augmented tag would have the form:

W	W_1	W_2	...	W_{k-1}	W_k	W
W'	W_1'	W_2'	...	W_{k-1}'	W_k'	W'

where the primed W's indicate complements. With ends of tags always forming

25   perfectly matched duplexes, all mismatched words will be internal mismatches thereby reducing the stability of tag-complement duplexes that otherwise would have mismatched words at their ends. It is well known that duplexes with internal mismatches are significantly less stable than duplexes with the same mismatch at a terminus.

30   A preferred embodiment of minimally cross-hybridizing sets are those whose subunits are made up of three of the four natural nucleotides. As will be discussed more fully below, the absence of one type of nucleotide in the oligonucleotide tags permits target polynucleotides to be loaded onto solid phase supports by use of the $5' \rightarrow 3'$ exonuclease activity of a DNA polymerase. The following is an exemplary

35   minimally cross-hybridizing set of subunits each comprising four nucleotides selected from the group consisting of A, G, and T:

5   Table II

Word:	w_1	w_2	w_3	w_4
Sequence:	GATT	TGAT	TAGA	TTTG

Word:	w_5	w_6	w_7	w_8
Sequence:	GTAA	AGTA	ATGT	AAAG

In this set, each member would form a duplex having three mismatched bases with
10   the complement of every other member.

Further exemplary minimally cross-hybridizing sets are listed below in Table
III. Clearly, additional sets can be generated by substituting different groups of
nucleotides, or by using subsets of known minimally cross-hybridizing sets.

15

Table III

Exemplary Minimally Cross-Hybridizing Sets of 4-mer Subunits

Set 1	Set 2	Set 3	Set 4	Set 5	Set 6
CATT	ACCC	AAAC	AAAG	AACA	AACG
CTAA	AGGG	ACCA	ACCA	ACAC	ACAA
TCAT	CACG	AGGG	AGGC	AGGG	AGGC
ACTA	CCGA	CACG	CACC	CAAG	CAAC
TACA	CGAC	CCGC	CCGG	CCGC	CCGG
TTTC	GAGC	CGAA	CGAA	CGCA	CGCA
ATCT	GCAG	GAGA	GAGA	GAGA	GAGA
AAAC	GGCA	GCAG	GCAC	GCCG	GCCC
	AAAA	GGCC	GGCG	GGAC	GGAG

Set 7	Set 8	Set 9	Set 10	Set 11	Set 12
AAGA	AAGC	AAGG	ACAG	ACCG	ACGA
ACAC	ACAA	ACAA	AACA	AAAA	AAAC
AGCG	AGCG	AGCC	AGGC	AGGC	AGCG
CAAG	CAAG	CAAC	CAAC	CACC	CACA
CCCA	CCCC	CCCG	CCGA	CCGA	CCAG
CGGC	CGGA	CGGA	CGCG	CGAG	CGGC
GACC	GACA	GACA	GAGG	GAGG	GAGG
GCGG	GCGG	GCGC	GCCC	GCAC	GCCC
GGAA	GGAC	GGAG	GGAA	GGCA	GGAA

The oligonucleotide tags of the invention and their complements are conveniently synthesized on an automated DNA synthesizer, e.g. an Applied Biosystems, Inc. (Foster City, California) model 392 or 394 DNA/RNA Synthesizer, using standard chemistries, such as phosphoramidite chemistry, e.g. disclosed in the following references: Beaucage and Iyer, Tetrahedron, 48: 2223-2311 (1992); Molko et al, U.S. patent 4,980,460; Koster et al, U.S. patent 4,725,677; Caruthers et al, U.S. patents 4,415,732; 4,458,066; and 4,973,679; and the like. Alternative chemistries, e.g. resulting in non-natural backbone groups, such as phosphorothioate, phosphoramidate, and the like, may also be employed provided that the resulting oligonucleotides are capable of specific hybridization. In some embodiments, tags may comprise naturally occurring nucleotides that permit processing or manipulation by enzymes, while the corresponding tag complements may comprise non-natural nucleotide analogs, such as peptide nucleic acids, or like compounds, that promote the formation of more stable duplexes during sorting.

When microparticles are used as supports, repertoires of oligonucleotide tags and tag complements may be generated by subunit-wise synthesis via "split and mix" techniques, e.g. as disclosed in Shortle et al, International patent application PCT/US93/03418 or Lyttle et al, Biotechniques, 19: 274-280 (1995). Briefly, the basic unit of the synthesis is a subunit of the oligonucleotide tag. Preferably, phosphoramidite chemistry is used and 3' phosphoramidite oligonucleotides are prepared for each subunit in a minimally cross-hybridizing set, e.g. for the set first listed above, there would be eight 4-mer 3'-phosphoramidites. Synthesis proceeds as disclosed by Shortle et al or in direct analogy with the techniques employed to generate diverse oligonucleotide libraries using nucleosidic monomers, e.g. as disclosed in Telenius et al, Genomics, 13: 718-725 (1992); Welsh et al, Nucleic Acids Research, 19: 5275-5279 (1991); Grothues et al, Nucleic Acids Research, 21: 1321-1322 (1993); Hartley, European patent application 90304496.4; Lam et al, Nature, 354: 82-84 (1991); Zuckerman et al, Int. J. Pept. Protein Research, 40: 498-507 (1992); and the like. Generally, these techniques simply call for the application of

mixtures of the activated monomers to the growing oligonucleotide during the coupling steps. Preferably, oligonucleotide tags and tag complements are synthesized on a DNA synthesizer having a number of synthesis chambers which is greater than or equal to the number of different kinds of words used in the construction of the tags.

5   That is, preferably there is a synthesis chamber corresponding to each type of word. In this embodiment, words are added nucleotide-by-nucleotide, such that if a word consists of five nucleotides there are five monomer couplings in each synthesis chamber. After a word is completely synthesized, the synthesis supports are removed from the chambers, mixed, and redistributed back to the chambers for the next cycle

10   of word addition. This latter embodiment takes advantage of the high coupling yields of monomer addition, e.g. in phosphoramidite chemistries.

Double stranded forms of tags may be made by separately synthesizing the complementary strands followed by mixing under conditions that permit duplex formation. Alternatively, double stranded tags may be formed by first synthesizing a

15   single stranded repertoire linked to a known oligonucleotide sequence that serves as a primer binding site. The second strand is then synthesized by combining the single stranded repertoire with a primer and extending with a polymerase. This latter approach is described in Oliphant et al, Gene, 44: 177-183 (1986). Such duplex tags may then be inserted into cloning vectors along with target polynucleotides for sorting

20   and manipulation of the target polynucleotide in accordance with the invention.

When tag complements are employed that are made up of nucleotides that have enhanced binding characteristics, such as PNAs or oligonucleotide $N3' \rightarrow P5'$ phosphoramidates, sorting can be implemented through the formation of D-loops between tags comprising natural nucleotides and their PNA or phosphoramidate

25   complements, as an alternative to the "stripping" reaction employing the $3' \rightarrow 5'$ exonuclease activity of a DNA polymerase to render a tag single stranded.

Oligonucleotide tags of the invention may range in length from 12 to 60 nucleotides or basepairs. Preferably, oligonucleotide tags range in length from 18 to 40 nucleotides or basepairs. More preferably, oligonucleotide tags range in length

30   from 25 to 40 nucleotides or basepairs. In terms of preferred and more preferred numbers of subunits, these ranges may be expressed as follows:

Table IV

Numbers of Subunits in Tags in Preferred Embodiments

35

Monomers in Subunit	Nucleotides in Oligonucleotide Tag		
	(12-60)	(18-40)	(25-40)

3	4-20 subunits	6-13 subunits	8-13 subunits
4	3-15 subunits	4-10 subunits	6-10 subunits
5	2-12 subunits	3-8 subunits	5-8 subunits
6	2-10 subunits	3-6 subunits	4-6 subunits

Most preferably, oligonucleotide tags are single stranded and specific hybridization occurs via Watson-Crick pairing with a tag complement.

Preferably, repertoires of single stranded oligonucleotide tags of the invention
5   contain at least 100 members; more preferably, repertoires of such tags contain at least 1000 members; and most preferably, repertoires of such tags contain at least 10,000 members.

Triplex Tags

In embodiments where specific hybridization occurs via triplex formation,
10   coding of tag sequences follows the same principles as for duplex-forming tags; however, there are further constraints on the selection of subunit sequences. Generally, third strand association via Hoogsteen type of binding is most stable along homopyrimidine-homopurine tracks in a double stranded target. Usually, base triplets form in T-A*T or C-G*C motifs (where "-" indicates Watson-Crick pairing and "*"
15   indicates Hoogsteen type of binding); however, other motifs are also possible. For example, Hoogsteen base pairing permits parallel and antiparallel orientations between the third strand (the Hoogsteen strand) and the purine-rich strand of the duplex to which the third strand binds, depending on conditions and the composition of the strands. There is extensive guidance in the literature for selecting appropriate
20   sequences, orientation, conditions, nucleoside type (e.g. whether ribose or deoxyribose nucleosides are employed), base modifications (e.g. methylated cytosine, and the like) in order to maximize, or otherwise regulate, triplex stability as desired in particular embodiments, e.g. Roberts et al, Proc. Natl. Acad. Sci., 88: 9397-9401 (1991); Roberts et al, Science, 258: 1463-1466 (1992); Roberts et al, Proc. Natl.
25   Acad. Sci., 93: 4320-4325 (1996); Distefano et al, Proc. Natl. Acad. Sci., 90: 1179-1183 (1993); Mergny et al, Biochemistry, 30: 9791-9798 (1991); Cheng et al, J. Am. Chem. Soc., 114: 4465-4474 (1992); Beal and Dervan, Nucleic Acids Research, 20: 2773-2776 (1992); Beal and Dervan, J. Am. Chem. Soc., 114: 4976-4982 (1992); Giovannangeli et al, Proc. Natl. Acad. Sci., 89: 8631-8635 (1992); Moser and Dervan,
30   Science, 238: 645-650 (1987); McShan et al, J. Biol. Chem., 267:5712-5721 (1992); Yoon et al, Proc. Natl. Acad. Sci., 89: 3840-3844 (1992); Blume et al, Nucleic Acids Research, 20: 1777-1784 (1992); Thuong and Helene, Angew. Chem. Int. Ed. Engl.

32: 666-690 (1993); Escude et al, Proc. Natl. Acad. Sci., 93: 4365-4369 (1996); and
the like. Conditions for annealing single-stranded or duplex tags to their single-
stranded or duplex complements are well known, e.g. Ji et al, Anal. Chem. 65: 1323-
1328 (1993); Cantor et al, U.S. patent 5,482,836; and the like. Use of triplex tags has
the advantage of not requiring a "stripping" reaction with polymerase to expose the
tag for annealing to its complement.

Preferably, oligonucleotide tags of the invention employing triplex
hybridization are double stranded DNA and the corresponding tag complements are
single stranded. More preferably, 5-methylcytosine is used in place of cytosine in the
tag complements in order to broaden the range of pH stability of the triplex formed
between a tag and its complement. Preferred conditions for forming triplexes are
fully disclosed in the above references. Briefly, hybridization takes place in
concentrated salt solution, e.g. 1.0 M NaCl, 1.0 M potassium acetate, or the like, at
pH below 5.5 (or 6.5 if 5-methylcytosine is employed). Hybridization temperature
depends on the length and composition of the tag; however, for an 18-20-mer tag of
longer, hybridization at room temperature is adequate. Washes may be conducted
with less concentrated salt solutions, e.g. 10 mM sodium acetate, 100 mM MgCl$_2$, pH
5.8, at room temperature. Tags may be eluted from their tag complements by
- incubation in a similar salt solution at pH 9.0.

Minimally cross-hybridizing sets of oligonucleotide tags that form triplexes
may be generated by the computer program of Appendix Ic, or similar programs. An
exemplary set of double stranded 8-mer words are listed below in capital letters with
the corresponding complements in small letters. Each such word differs from each of
the other words in the set by three base pairs.

<div align="center">

Table V

Exemplary Minimally Cross-Hybridizing
Set of DoubleStranded 8-mer Tags

</div>

5'-AAGGAGAG	5'-AAAGGGGA	5'-AGAGAAGA	5'-AGGGGGGG
3'-TTCCTCTC	3'-TTTCCCCT	3'-TCTCTTCT	3'-TCCCCCCC
3'-ttcctctc	3'-tttccccct	3'-tctcttct	3'-tccccccc
5'-AAAAAAAA	5'-AAGAGAGA	5'-AGGAAAAG	5'-GAAAGGAG
3'-TTTTTTTT	3'-TTCTCTCT	3'-TCCTTTTC	3'-CTTTCCTC
3'-tttttttt	3'-ttctctct	3'-tccttttc	3'-ctttcctc
5'-AAAAAGGG	5'-AGAAGAGG	5'-AGGAAGGA	5'-GAAGAAGG
3'-TTTTTCCC	3'-TCTTCTCC	3'-TCCTTCCT	3'-CTTCTTCC
3'-tttttccc	3'-tcttctcc	3'-tccttcct	3'-cttcttcc
5'-AAAGGAAG	5'-AGAAGGAA	5'-AGGGGAAA	5'-GAAGAGAA
3'-TTTCCTTC	3'-TCTTCCTT	3'-TCCCCTTT	3'-CTTCTCTT
3'-tttccttc	3'-tcttcctt	3'-tcccccttt	3'-cttctctt

5

10

Table VI
Repertoire Size of Various Double Stranded Tags
That Form Triplexes with Their Tag Complements

Oligonucleotide Word Length	Nucleotide Difference between Oligonucleotides of Minimally Cross-Hybridizing Set	Maximal Size of Minimally Cross-Hybridizing Set	Size of Repertoire with Four Words	Size of Repertoire with Five Words
4	2	8	4096	3.2×10^4
6	3	8	4096	3.2×10^4
8	3	16	6.5×10^4	1.05×10^6
10	5	8	4096	
15	5	92		
20	6	765		
20	8	92		
20	10	22		

15   Preferably, repertoires of double stranded oligonucleotide tags of the invention contain at least 10 members; more preferably, repertoires of such tags contain at least 100 members. Preferably, words are between 4 and 8 nucleotides in length for combinatorially synthesized double stranded oligonucletide tags, and oligonucleotide tags are between 12 and 60 base pairs in length. More preferably, such tags are

20   between 18 and 40 base pairs in length.

Solid Phase Supports

     Solid phase supports for use with the invention may have a wide variety of forms, including microparticles, beads, and membranes, slides, plates, micromachined

25   chips, and the like. Likewise, solid phase supports of the invention may comprise a

wide variety of compositions, including glass, plastic, silicon, alkanethiolate-derivatized gold, cellulose, low cross-linked and high cross-linked polystyrene, silica gel, polyamide, and the like. Preferably, either a population of discrete particles are employed such that each has a uniform coating, or population, of complementary

5   sequences of the same tag (and no other), or a single or a few supports are employed with spatially discrete regions each containing a uniform coating, or population, of complementary sequences to the same tag (and no other). In the latter embodiment, the area of the regions may vary according to particular applications; usually, the regions range in area from several μm^2, e.g. 3-5, to several hundred μm^2, e.g. 100-

10   500. Preferably, such regions are spatially discrete so that signals generated by events, e.g. fluorescent emissions, at adjacent regions can be resolved by the detection system being employed. In some applications, it may be desirable to have regions with uniform coatings of more than one tag complement, e.g. for simultaneous sequence analysis, or for bringing separately tagged molecules into close proximity.

15   Tag complements may be used with the solid phase support that they are synthesized on, or they may be separately synthesized and attached to a solid phase support for use, e.g. as disclosed by Lund et al, Nucleic Acids Research, 16: 10861-10880 (1988); Albretsen et al, Anal. Biochem., 189: 40-50 (1990); Wolf et al, Nucleic Acids Research, 15: 2911-2926 (1987); or Ghosh et al, Nucleic Acids Research, 15:

20   5353-5372 (1987). Preferably, tag complements are synthesized on and used with the same solid phase support, which may comprise a variety of forms and include a variety of linking moieties. Such supports may comprise microparticles or arrays, or matrices, of regions where uniform populations of tag complements are synthesized. A wide variety of microparticle supports may be used with the invention, including

25   microparticles made of controlled pore glass (CPG), highly cross-linked polystyrene, acrylic copolymers, cellulose, nylon, dextran, latex, polyacrolein, and the like, disclosed in the following exemplary references: Meth. Enzymol., Section A, pages 11-147, vol. 44 (Academic Press, New York, 1976); U.S. patents 4,678,814; 4,413,070; and 4,046;720; and Pon, Chapter 19, in Agrawal, editor, Methods in

30   Molecular Biology, Vol. 20, (Humana Press, Totowa, NJ, 1993). Microparticle supports further include commercially available nucleoside-derivatized CPG and polystyrene beads (e.g. available from Applied Biosystems, Foster City, CA); derivatized magnetic beads; polystyrene grafted with polyethylene glycol (e.g., TentaGel^{TM}, Rapp Polymere, Tubingen Germany); and the like. Selection of the

35   support characteristics, such as material, porosity, size, shape, and the like, and the type of linking moiety employed depends on the conditions under which the tags are used. For example, in applications involving successive processing with enzymes, supports and linkers that minimize steric hindrance of the enzymes and that facilitate

access to substrate are preferred. Other important factors to be considered in selecting the most appropriate microparticle support include size uniformity, efficiency as a synthesis support, degree to which surface area known, and optical properties, e.g. as explain more fully below, clear smooth beads provide instrumentational advantages
5   when handling large numbers of beads on a surface.

Exemplary linking moieties for attaching and/or synthesizing tags on microparticle surfaces are disclosed in Pon et al, Biotechniques, 6:768-775 (1988); Webb, U.S. patent 4,659,774; Barany et al, International patent application PCT/US91/06103; Brown et al, J. Chem. Soc. Commun., 1989: 891-893; Damha et
10   al, Nucleic Acids Research, 18: 3813-3821 (1990); Beattie et al, Clinical Chemistry, 39: 719-722 (1993); Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992); and the like.

As mentioned above, tag complements may also be synthesized on a single (or a few) solid phase support to form an array of regions uniformly coated with tag
15   complements. That is, within each region in such an array the same tag complement is synthesized. Techniques for synthesizing such arrays are disclosed in McGall et al, International application PCT/US93/03767; Pease et al, Proc. Natl. Acad. Sci., 91: 5022-5026 (1994); Southern and Maskos, International application -PCT/GB89/01114; Maskos and Southern (cited above); Southern et al, Genomics, 13:
20   1008-1017 (1992); and Maskos and Southern, Nucleic Acids Research, 21: 4663-4669 (1993).

Preferably, the invention is implemented with microparticles or beads uniformly coated with complements of the same tag sequence. Microparticle supports and methods of covalently or noncovalently linking oligonucleotides to their surfaces
25   are well known, as exemplified by the following references: Beaucage and Iyer (cited above); Gait, editor, Oligonucleotide Synthesis: A Practical Approach (IRL Press, Oxford, 1984); and the references cited above. Generally, the size and shape of a microparticle is not critical; however, microparticles in the size range of a few, e.g. 1-2, to several hundred, e.g. 200-1000 μm diameter are preferable, as they facilitate the
30   construction and manipulation of large repertoires of oligonucleotide tags with minimal reagent and sample usage.

In some preferred applications, commercially available controlled-pore glass (CPG) or polystyrene supports are employed as solid phase supports in the invention. Such supports come available with base-labile linkers and initial nucleosides attached,
35   e.g. Applied Biosystems (Foster City, CA). Preferably, microparticles having pore size between 500 and 1000 angstroms are employed.

In other preferred applications, non-porous microparticles are employed for their optical properties, which may be advantageously used when tracking large

numbers of microparticles on planar supports, such as a microscope slide. Particularly preferred non-porous microparticles are the glycidal methacrylate (GMA) beads available from Bangs Laboratories (Carmel, IN). Such microparticles are useful in a variety of sizes and derivatized with a variety of linkage groups for

5   synthesizing tags or tag complements. Preferably, for massively parallel manipulations of tagged microparticles, 5 μm diameter GMA beads are employed.

10

Attaching Tags to Polynucleotides
For Sorting onto Solid Phase Supports

An important aspect of the invention is the sorting and attachment of a populations of polynucleotides, e.g. from a cDNA library, to microparticles or to

15   separate regions on a solid phase support such that each microparticle or region has substantially only one kind of polynucleotide attached. This objective is accomplished by insuring that substantially all different polynucleotides have different tags attached. This condition, in turn, is brought about by taking a sample of the full ensemble of tag-polynucleotide conjugates for analysis. (It is acceptable that

20   identical polynucleotides have different tags, as it merely results in the same polynucleotide being operated on or analyzed twice in two different locations.) Such sampling can be carried out either overtly--for example, by taking a small volume from a larger mixture--after the tags have been attached to the polynucleotides, it can be carried out inherently as a secondary effect of the techniques used to process the

25   polynucleotides and tags, or sampling can be carried out both overtly and as an inherent part of processing steps.

Preferably, in constructing a cDNA library where substantially all different cDNAs have different tags, a tag repertoire is employed whose complexity, or number of distinct tags, greatly exceeds the total number of mRNAs extracted from a cell or

30   tissue sample. Preferably, the complexity of the tag repertoire is at least 10 times that of the polynucleotide population; and more preferably. the complexity of the tag repertoire is at least 100 times that of the polynucleotide population. Below, a protocol is disclosed for cDNA library construction using a primer mixture that contains a full repertoire of exemplary 9-word tags. Such a mixture of tag-containing

35   primers has a complexity of 8^9, or about 1.34×10^8. As indicated by Winslow et al, Nucleic Acids Research, 19: 3251-3253 (1991), mRNA for library construction can be extracted from as few as 10-100 mammalian cells. Since a single mammalian cell contains about 5×10^5 copies of mRNA molecules of about 3.4×10^4 different kinds,

by standard techniques one can isolate the mRNA from about 100 cells, or (theoretically) about 5×10^7 mRNA molecules. Comparing this number to the complexity of the primer mixture shows that without any additional steps, and even assuming that mRNAs are converted into cDNAs with perfect efficiency (1% efficiency or less is more accurate), the cDNA library construction protocol results in a population containing no more than 37% of the total number of different tags. That is, without any overt sampling step at all, the protocol inherently generates a sample that comprises 37%, or less, of the tag repertoire. The probability of obtaining a double under these conditions is about 5%, which is within the preferred range. With mRNA from 10 cells, the fraction of the tag repertoire sampled is reduced to only 3.7%, even assuming that all the processing steps take place at 100% efficiency. In fact, the efficiencies of the processing steps for constructing cDNA libraries are very low, a "rule of thumb" being that good library should contain about 10^8 cDNA clones from mRNA extracted from 10^6 mammalian cells.

Use of larger amounts of mRNA in the above protocol, or for larger amounts of polynucleotides in general, where the number of such molecules exceeds the complexity of the tag repertoire, a tag-polynucleotide conjugate mixture potentially contains every possible pairing of tags and types of mRNA or polynucleotide. In such cases, overt sampling may be implemented by removing a sample volume after a serial dilution of the starting mixture of tag-polynucleotide conjugates. The amount of dilution required depends on the amount of starting material and the efficiencies of the processing steps, which are readily estimated.

If mRNA were extracted from 10^6 cells (which would correspond to about 0.5 µg of poly(A)$^+$ RNA), and if primers were present in about 10-100 fold concentration excess--as is called for in a typical protocol, e.g. Sambrook et al, Molecular Cloning, Second Edition, page 8.61 [10 µL 1.8 kb mRNA at 1 mg/mL equals about 1.68×10^{-11} moles and 10 µL 18-mer primer at 1 mg/mL equals about 1.68×10^{-9} moles], then the total number of tag-polynucleotide conjugates in a cDNA library would simply be equal to or less than the starting number of mRNAs, or about 5×10^{11} vectors containing tag-polynucleotide conjugates--again this assumes that each step in cDNA construction--first strand synthesis, second strand synthesis, ligation into a vector-- occurs with perfect efficiency, which is a very conservative estimate. The actual number is significantly less.

If a sample of n tag-polynucleotide conjugates are randomly drawn from a reaction mixture--as could be effected by taking a sample volume, the probability of drawing conjugates having the same tag is described by the Poisson distribution, $P(r) = e^{-\lambda}(\lambda)^r/r$, where r is the number of conjugates having the same tag and $\lambda = np$, where p is the probability of a given tag being selected. If $n = 10^6$ and $p = 1/(1.34 \times$

10^8), then $\lambda = .00746$ and $P(2) = 2.76 \times 10^{-5}$. Thus, a sample of one million molecules gives rise to an expected number of doubles well within the preferred range. Such a sample is readily obtained as follows: Assume that the 5×10^{11} mRNAs are perfectly converted into 5×10^{11} vectors with tag-cDNA conjugates as inserts and that the $5 \times$

5   10^{11} vectors are in a reaction solution having a volume of 100 μl. Four 10-fold serial dilutions may be carried out by transferring 10 μl from the original solution into a vessel containing 90 μl of an appropriate buffer, such as TE. This process may be repeated for three additional dilutions to obtain a 100 μl solution containing 5×10^5 vector molecules per μl. A 2 μl aliquot from this solution yields 10^6 vectors

10   containing tag-cDNA conjugates as inserts. This sample is then amplified by straight forward transformation of a competent host cell followed by culturing.

  Of course, as mentioned above, no step in the above process proceeds with perfect efficiency. In particular, when vectors are employed to amplify a sample of tag-polynucleotide conjugates, the step of transforming a host is very inefficient.

15   Usually, no more than 1% of the vectors are taken up by the host and replicated. Thus, for such a method of amplification, even fewer dilutions would be required to obtain a sample of 10^6 conjugates.

  A repertoire of oligonucleotide tags can be conjugated to a population of polynucleotides in a number of ways, including direct enzymatic ligation,

20   amplification, e.g. via PCR, using primers containing the tag sequences, and the like. The initial ligating step produces a very large population of tag-polynucleotide conjugates such that a single tag is generally attached to many different polynucleotides. However, as noted above, by taking a sufficiently small sample of the conjugates, the probability of obtaining "doubles," i.e. the same tag on two

25   different polynucleotides, can be made negligible. Generally, the larger the sample the greater the probability of obtaining a double. Thus, a design trade-off exists between selecting a large sample of tag-polynucleotide conjugates--which, for example, ensures adequate coverage of a target polynucleotide in a shotgun sequencing operation or adequate representation of a rapidly changing mRNA pool,

30   and selecting a small sample which ensures that a minimal number of doubles will be present. In most embodiments, the presence of doubles merely adds an additional source of noise or, in the case of sequencing, a minor complication in scanning and signal processing, as microparticles giving multiple fluorescent signals can simply be ignored.

35   As used herein, the term "substantially all" in reference to attaching tags to molecules, especially polynucleotides, is meant to reflect the statistical nature of the sampling procedure employed to obtain a population of tag-molecule conjugates essentially free of doubles. The meaning of substantially all in terms of actual

percentages of tag-molecule conjugates depends on how the tags are being employed.
Preferably, for nucleic acid sequencing, substantially all means that at least eighty
percent of the polynucleotides have unique tags attached. More preferably, it means
that at least ninety percent of the polynucleotides have unique tags attached. Still
5   more preferably, it means that at least ninety-five percent of the polynucleotides have
unique tags attached. And, most preferably, it means that at least ninety-nine percent
of the polynucleotides have unique tags attached.

        Preferably, when the population of polynucleotides consists of messenger
RNA (mRNA), oligonucleotides tags may be attached by reverse transcribing the
10   mRNA with a set of primers preferably containing complements of tag sequences.
An exemplary set of such primers could have the following sequence (SEQ ID NO:
1):

              5'-mRNA- [A]$_n$ -3'
15   [T]$_{19}$GG[W,W,W,C]$_9$AC̲C̲A̲G̲C̲T̲GATC-5'-biotin

        where "[W,W,W,C]9" represents the sequence of an oligonucleotide tag of nine
        subunits of four nucleotides each and "[W,W,W,C]" represents the subunit sequences
20   listed above, i.e. "W" represents T or A. The underlined sequences identify an
optional restriction endonuclease site that can be used to release the polynucleotide
from attachment to a solid phase support via the biotin, if one is employed. For the
above primer, the complement attached to a microparticle could have the form:

25   5'-[G,W,W,W]$_9$TGG-linker-microparticle

        After reverse transcription, the mRNA is removed, e.g. by RNase H digestion.
and the second strand of the cDNA is synthesized using, for example, a primer of the
following form (SEQ ID NO: 2):
30

                5'-NRRGATCYNNN-3'

where N is any one of A, T, G, or C; R is a purine-containing nucleotide, and Y is a
pyrimidine-containing nucleotide. This particular primer creates a Bst Y1 restriction
35   site in the resulting double stranded DNA which, together with the Sal I site,
facilitates cloning into a vector with, for example, Bam HI and Xho I sites. After Bst
Y1 and Sal I digestion, the exemplary conjugate would have the form:

$$5' - \text{RCGACCA}[C,W,W,W]_9\text{GG}[T]_{19} - \text{cDNA} - \text{NNNR}$$
$$\text{GGT}[G,W,W,W]_9\text{CC}[A]_{19} - \text{rDNA} - \text{NNNYCTAG} - 5'$$

The polynucleotide-tag conjugates may then be manipulated using standard molecular

5　biology techniques. For example, the above conjugate--which is actually a mixture-- may be inserted into commercially available cloning vectors, e.g. Stratagene Cloning System (La Jolla, CA); transfected into a host, such as a commercially available host bacteria; which is then cultured to increase the number of conjugates. The cloning vectors may then be isolated using standard techniques, e.g. Sambrook et al,

10　Molecular Cloning, Second Edition (Cold Spring Harbor Laboratory, New York, 1989). Alternatively, appropriate adaptors and primers may be employed so that the conjugate population can be increased by PCR.

　　　　Preferably, when the ligase-based method of sequencing is employed, the Bst Y1 and Sal I digested fragments are cloned into a Bam HI-/Xho I-digested vector

15　having the following single-copy restriction sites (SEQ ID NO: 3):

$$5' - \text{GAGGATGCCTTTATGGATCCACTCGAGATCCCAATCCA} - 3'$$
$$\phantom{5' - \text{GA}}\text{FokI}\text{BamHI}\text{XhoI}$$

20

This adds the Fok I site which will allow initiation of the sequencing process discussed more fully below.

　　　　Tags can be conjugated to cDNAs of existing libraries by standard cloning methods. cDNAs are excised from their existing vector, isolated, and then ligated into

25　a vector containing a repertoire of tags. Preferably, the tag-containing vector is linearized by cleaving with two restriction enzymes so that the excised cDNAs can be ligated in a predetermined orientation. The concentration of the linearized tag- containing vector is in substantial excess over that of the cDNA inserts so that ligation provides an inherent sampling of tags.

30　　　　A general method for exposing the single stranded tag after amplification involves digesting a target polynucleotide-containing conjugate with the $5' {\rightarrow} 3'$ exonuclease activity of T4 DNA polymerase, or a like enzyme. When used in the presence of a single deoxynucleoside triphosphate, such a polymerase will cleave nucleotides from 3' recessed ends present on the non-template strand of a double

35　stranded fragment until a complement of the single deoxynucleoside triphosphate is reached on the template strand. When such a nucleotide is reached the $5' {\rightarrow} 3'$ digestion effectively ceases, as the polymerase's extension activity adds nucleotides at a higher rate than the excision activity removes nucleotides. Consequently, single

stranded tags constructed with three nucleotides are readily prepared for loading onto solid phase supports.

The technique may also be used to preferentially methylate interior Fok I sites of a target polynucleotide while leaving a single Fok I site at the terminus of the polynucleotide unmethylated. First, the terminal Fok I site is rendered single stranded using a polymerase with deoxycytidine triphosphate. The double stranded portion of the fragment is then methylated, after which the single stranded terminus is filled in with a DNA polymerase in the presence of all four nucleoside triphosphates, thereby regenerating the Fok I site. Clearly, this procedure can be generalized to endonucleases other than Fok I.

After the oligonucleotide tags are prepared for specific hybridization, e.g. by rendering them single stranded as described above, the polynucleotides are mixed with microparticles containing the complementary sequences of the tags under conditions that favor the formation of perfectly matched duplexes between the tags and their complements. There is extensive guidance in the literature for creating these conditions. Exemplary references providing such guidance include Wetmur, Critical Reviews in Biochemistry and Molecular Biology, 26: 227-259 (1991); Sambrook et al, Molecular Cloning: A Laboratory Manual, 2nd Edition (Cold Spring Harbor Laboratory, New York, 1989); and the like. Preferably, the hybridization conditions are sufficiently stringent so that only perfectly matched sequences form stable duplexes. Under such conditions the polynucleotides specifically hybridized through their tags may be ligated to the complementary sequences attached to the microparticles. Finally, the microparticles are washed to remove polynucleotides with unligated and/or mismatched tags.

When CPG microparticles conventionally employed as synthesis supports are used, the density of tag complements on the microparticle surface is typically greater than that necessary for some sequencing operations. That is, in sequencing approaches that require successive treatment of the attached polynucleotides with a variety of enzymes, densely spaced polynucleotides may tend to inhibit access of the relatively bulky enzymes to the polynucleotides. In such cases, the polynucleotides are preferably mixed with the microparticles so that tag complements are present in significant excess, e.g. from 10:1 to 100:1, or greater, over the polynucleotides. This ensures that the density of polynucleotides on the microparticle surface will not be so high as to inhibit enzyme access. Preferably, the average inter-polynucleotide spacing on the microparticle surface is on the order of 30-100 nm. Guidance in selecting ratios for standard CPG supports and Ballotini beads (a type of solid glass support) is found in Maskos and Southern, Nucleic Acids Research, 20: 1679-1684 (1992). Preferably, for sequencing applications, standard CPG beads of diameter in the range

of 20-50 μm are loaded with about 10^5 polynucleotides, and GMA beads of diameter in the range of 5-10 μm are loaded with a few tens of thousand of polynucleotides, e.g. 4×10^4 to 6×10^4.

5   In the preferred embodiment, tag complements are synthesized on microparticles combinatorially; thus, at the end of the synthesis, one obtains a complex mixture of microparticles from which a sample is taken for loading tagged polynucleotides. The size of the sample of microparticles will depend on several factors, including the size of the repertoire of tag complements, the nature of the apparatus for used for observing loaded microparticles--e.g. its capacity, the tolerance

10   for multiple copies of microparticles with the same tag complement (i.e. "bead doubles"), and the like. The following table provide guidance regarding microparticle sample size, microparticle diameter, and the approximate physical dimensions of a packed array of microparticles of various diameters.

15

Microparticle diameter	5 μm	10 μm	20 μm	40 μm
Max. no. polynucleotides loaded at 1 per 10^5 sq. angstrom		3×10^5	1.26×10^6	5×10^6
Approx. area of monolayer of 10^6 microparticles	.45 x .45 cm	1 x 1 cm	2 x 2 cm	4 x 4 cm

20   The probability that the sample of microparticles contains a given tag complement or is present in multiple copies is described by the Poisson distribution, as indicated in the following table.

25

Table VII

Number of microparticles in sample (as fraction of repertoire size), m	Fraction of repertoire of tag complements present in sample, $1-e^{-m}$	Fraction of microparticles in sample with unique tag complement attached, $m(e^{-m})/2$	Fraction of microparticles in sample carrying same tag complement as one other microparticle in sample ("bead doubles"), $m^2(e^{-m})/2$
1.000	0.63	0.37	0.18
.693	0.50	0.35	0.12
.405	0.33	0.27	0.05
.285	0.25	0.21	0.03
.223	0.20	0.18	0.02
.105	0.10	0.09	0.005
.010	0.01	0.01	

High Specificity Sorting and Panning

The kinetics of sorting depends on the rate of hybridization of oligonucleotide tags to their tag complements which, in turn, depends on the complexity of the tags in the hybridization reaction. Thus, a trade off exists between sorting rate and tag complexity, such that an increase in sorting rate may be achieved at the cost of reducing the complexity of the tags involved in the hybridization reaction. As explained below, the effects of this trade off may be ameliorated by "panning."

Specificity of the hybridizations may be increased by taking a sufficiently small sample so that both a high percentage of tags in the sample are unique and the nearest neighbors of substantially all the tags in a sample differ by at least two words. This latter condition may be met by taking a sample that contains a number of tag-polynucleotide conjugates that is about 0.1 percent or less of the size of the repertoire being employed. For example, if tags are constructed with eight words selected from Table II, a repertoire of 8^8, or about 1.67×10^7, tags and tag complements are produced. In a library of tag-cDNA conjugates as described above, a 0.1 percent sample means that about 16,700 different tags are present. If this were loaded directly onto a repertoire-equivalent of microparticles, or in this example a sample of 1.67×10^7 microparticles, then only a sparse subset of the sampled microparticles would be loaded. The density of loaded microparticles can be increase--for example, for more efficient sequencing--by undertaking a "panning" step in which the sampled tag-cDNA conjugates are used to separate loaded microparticles from unloaded microparticles. Thus, in the example above, even though a "0.1 percent" sample

contains only 16,700 cDNAs, the sampling and panning steps may be repeated until as many loaded microparticles as desired are accumulated.

A panning step may be implemented by providing a sample of tag-cDNA conjugates each of which contains a capture moiety at an end opposite, or distal to,
5   the oligonucleotide tag. Preferably, the capture moiety is of a type which can be released from the tag-cDNA conjugates, so that the tag-cDNA conjugates can be sequenced with a single-base sequencing method. Such moieties may comprise biotin, digoxigenin, or like ligands, a triplex binding region, or the like. Preferably, such a capture moiety comprises a biotin component. Biotin may be attached to tag-
10   cDNA conjugates by a number of standard techniques. If appropriate adapters containing PCR primer binding sites are attached to tag-cDNA conjugates, biotin may be attached by using a biotinylated primer in an amplification after sampling. Alternatively, if the tag-cDNA conjugates are inserts of cloning vectors, biotin may be attached after excising the tag-cDNA conjugates by digestion with an appropriate
15   restriction enzyme followed by isolation and filling in a protruding strand distal to the tags with a DNA polymerase in the presence of biotinylated uridine triphosphate.

After a tag-cDNA conjugate is captured, it may be released from the biotin moiety in a number of ways, such as by a chemical linkage that is cleaved by -reduction, e.g. Herman et al, Anal. Biochem., 156: 48-55 (1986), or that is cleaved
20   photochemically, e.g. Olejnik et al, Nucleic Acids Research, 24: 361-366 (1996), or that is cleaved enzymatically by introducing a restriction site in the PCR primer. The latter embodiment can be exemplified by considering the library of tag-polynucleotide conjugates described above:

25   5'-RCGACCA[C,W,W,W]$_9$GG[T]$_{19}$- cDNA -NNNR
                GGT[G,W,W,W]$_9$CC[A]$_{19}$- rDNA -NNNYCTAG-5'

The following adapters may be ligated to the ends of these fragments to permit amplification by PCR:
30

              5'- XXXXXXXXXXXXXXXXXXXX
                 XXXXXXXXXXXXXXXXXXXXYGAT

35   Right Adapter

              GATCZZ**ACTAGT**ZZZZZZZZZZZZ-3'
40   ZZ**TGATCA**ZZZZZZZZZZZZ

Left Adapter

ZZ**TGATCA**ZZZZZZZZZZZZ-5'-biotin

Left Primer

where "ACTAGT" is a Spe I recognition site (which leaves a staggered cleavage
ready for single base sequencing), and the X's and Z's are nucleotides selected so that
the annealing and dissociation temperatures of the respective primers are
approximately the same. After ligation of the adapters and amplification by PCR
using the biotinylated primer, the tags of the conjugates are rendered single stranded
by the exonuclease activity of T4 DNA polymerase and conjugates are combined with
a sample of microparticles; e.g. a repertoire equivalent, with tag complements
attached. After annealing under stringent conditions (to minimize mis-attachment of
tags), the conjugates are preferably ligated to their tag complements and the loaded
microparticles are separated from the unloaded microparticles by capture with
avidinated magnetic beads, or like capture technique.

Returning to the example, this process results in the accumulation of about
10,500 (=16,700 x .63) loaded microparticles with different tags, which may be
released from the magnetic beads by cleavage with Spe I. By repeating this process
40-50 times with new samples of microparticles and tag-cDNA conjugates, 4-5 x 10^5
cDNAs can be accumulated by pooling the released microparticles. The pooled
microparticles may then be simultaneously sequenced by a single-base sequencing
technique.

Determining how many times to repeat the sampling and panning steps--or
more generally, determining how many cDNAs to analyze, depends on one's
objective. If the objective is to monitor the changes in abundance of relatively
common sequences; e.g. making up 5% or more of a population, then relatively small
samples, i.e. a small fraction of the total population size, may allow statistically
significant estimates of relative abundances. On the other hand, if one seeks to
monitor the abundances of rare sequences, e.g. making up 0.1% or less of a
population, then large samples are required. Generally, there is a direct relationship
between sample size and the reliability of the estimates of relative abundances based
on the sample. There is extensive guidance in the literature on determining
appropriate sample sizes for making reliable statistical estimates, e.g. Koller et al,
Nucleic Acids Research, 23:185-191 (1994); Good, Biometrika, 40: 16-264 (1953);
Bunge et al, J. Am. Stat. Assoc., 88: 364-373 (1993); and the like. Preferably, for

monitoring changes in gene expression based on the analysis of a series of cDNA libraries containing 10^5 to 10^8 independent clones of 3.0-3.5×10^4 different sequences, a sample of at least 10^4 sequences are accumulated for analysis of each library. More preferably, a sample of at least 10^5 sequences are accumulated for the

5   analysis of each library; and most preferably, a sample of at least 5×10^5 sequences are accumulated for the analysis of each library. Alternatively, the number of sequences sampled is preferably sufficient to estimate the relative abundance of a sequence present at a frequency within the range of 0.1% to 5% with a 95% confidence limit no larger than 0.1% of the population size.

10

Single Base DNA Sequencing

The present invention can be employed with conventional methods of DNA sequencing, e.g. as disclosed by Hultman et al, Nucleic Acids Research, 17: 4937-4946 (1989). However, for parallel, or simultaneous, sequencing of multiple

15   polynucleotides, a DNA sequencing methodology is preferred that requires neither electrophoretic separation of closely sized DNA fragments nor analysis of cleaved nucleotides by a separate analytical procedure, as in peptide sequencing. Preferably, the methodology permits the stepwise identification of nucleotides, usually one at a - time, in a sequence through successive cycles of treatment and detection. Such

20   methodologies are referred to herein as "single base" sequencing methods. Single base approaches are disclosed in the following references: Cheeseman, U.S. patent 5,302,509; Tsien et al, International application WO 91/06678; Rosenthal et al, International application WO 93/21340; Canard et al, Gene, 148: 1-6 (1994); and Metzker et al, Nucleic Acids Research, 22: 4259-4267 (1994).

25   A "single base" method of DNA sequencing which is suitable for use with the present invention and which requires no electrophoretic separation of DNA fragments is described in International application PCT/US95/03678. Briefly, the method comprises the following steps: (a) ligating a probe to an end of the polynucleotide having a protruding strand to form a ligated complex, the probe having a

30   complementary protruding strand to that of the polynucleotide and the probe having a nuclease recognition site; (b) removing unligated probe from the ligated complex; (c) identifying one or more nucleotides in the protruding strand of the polynucleotide by the identity of the ligated probe; (d) cleaving the ligated complex with a nuclease; and (e) repeating steps (a) through (d) until the nucleotide sequence of the polynucleotide,

35   or a portion thereof, is determined.

A single signal generating moiety, such as a single fluorescent dye, may be employed when sequencing several different target polynucleotides attached to different spatially addressable solid phase supports, such as fixed microparticles, in a

parallel sequencing operation. This may be accomplished by providing four sets of
probes that are applied sequentially to the plurality of target polynucleotides on the
different microparticles. An exemplary set of such probes are shown below:

5

Set 1	Set 2	Set 3	Set 4
ANNNN...NN	dANNNN...NN	dANNNN...NN	dANNNN...NN
N...NNTT...T*	d N...NNTT...T	N...NNTT...T	N...NNTT...T
dCNNNN...NN	CNNNN...NN	dCNNNN...NN	dCNNNN...NN
N...NNTT...T	N...NNTT...T*	N...NNTT...T	N...NNTT...T
dGNNNN...NN	dGNNNN...NN	GNNNN...NN	dGNNNN...NN
N...NNTT...T	N...NNTT...T	N...NNTT...T*	N...NNTT...T
dTNNNN...NN	dTNNNN...NN	dTNNNN...NN	TNNNN...NN
N...NNTT...T	N...NNTT...T	N...NNTT...T	N...NNTT...T*

where each of the listed probes represents a mixture of $4^3=64$ oligonucleotides such
that the identity of the 3' terminal nucleotide of the top strand is fixed and the other
positions in the protruding strand are filled by every 3-mer permutation of nucleotides,
10　or complexity reducing analogs. The listed probes are also shown with a single
stranded poly-T tail with a signal generating moiety attached to the terminal thymidine,
shown as "T*". The "d" on the unlabeled probes designates a ligation-blocking moiety
or absense of 3'-hydroxyl, which prevents unlabeled probes from being ligated.
Preferably, such 3'-terminal nucleotides are dideoxynucleotides. In this embodiment,
15　the probes of set 1 are first applied to the plurality of target polynucleotides and treated
with a ligase so that target polynucleotides having a thymidine complementary to the 3'
terminal adenosine of the labeled probes are ligated. The unlabeled probes are
simultaneously applied to minimize inappropriate ligations. The locations of the target
polynucleotides that form ligated complexes with probes terminating in "A" are
20　identified by the signal generated by the label carried on the probe. After washing and
cleavage, the probes of set 2 are applied. In this case, target polynucleotides forming
ligated complexes with probes terminating in "C" are identified by location. Similarly,
the probes of sets 3 and 4 are applied and locations of positive signals identified. This
process of sequentially applying the four sets of probes continues until the desired
25　number of nucleotides are identified on the target polynucleotides. Clearly, one of
ordinary skill could construct similar sets of probes that could have many variations,
such as having protruding strands of different lengths, different moieties to block
ligation of unlabeled probes, different means for labeling probes, and the like.

Apparatus for Sequencing Populations of Polynucleotides

An objective of the invention is to sort identical molecules, particularly polynucleotides, onto the surfaces of microparticles by the specific hybridization of tags and their complements. Once such sorting has taken place, the presence of the

5   molecules or operations performed on them can be detected in a number of ways depending on the nature of the tagged molecule, whether microparticles are detected separately or in "batches," whether repeated measurements are desired, and the like. Typically, the sorted molecules are exposed to ligands for binding, e.g. in drug development, or are subjected chemical or enzymatic processes, e.g. in polynucleotide

10   sequencing. In both of these uses it is often desirable to simultaneously observe signals corresponding to such events or processes on large numbers of microparticles. Microparticles carrying sorted molecules (referred to herein as "loaded" microparticles) lend themselves to such large scale parallel operations, e.g. as demonstrated by Lam et al (cited above).

15   Preferably, whenever light-generating signals, e.g. chemiluminescent, fluorescent, or the like, are employed to detect events or processes, loaded microparticles are spread on a planar substrate, e.g. a glass slide, for examination with a scanning system, such as described in International patent applications
PCT/US91/09217, PCT/NL90/00081, and PCT/US95/01886. The scanning system

20   should be able to reproducibly scan the substrate and to define the positions of each microparticle in a predetermined region by way of a coordinate system. In polynucleotide sequencing applications, it is important that the positional identification of microparticles be repeatable in successive scan steps.

Such scanning systems may be constructed from commercially available

25   components, e.g. x-y translation table controlled by a digital computer used with a detection system comprising one or more photomultiplier tubes, or alternatively, a CCD array, and appropriate optics, e.g. for exciting, collecting, and sorting fluorescent signals. In some embodiments a confocal optical system may be desirable. An exemplary scanning system suitable for use in four-color sequencing is

30   illustrated diagrammatically in Figure 5. Substrate 300, e.g. a microscope slide with fixed microparticles, is placed on x-y translation table 302, which is connected to and controlled by an appropriately programmed digital computer 304 which may be any of a variety of commercially available personal computers, e.g. 486-based machines or PowerPC model 7100 or 8100 available form Apple Computer (Cupertino, CA).

35   Computer software for table translation and data collection functions can be provided by commercially available laboratory software, such as Lab Windows, available from National Instruments.

Substrate 300 and table 302 are operationally associated with microscope 306 having one or more objective lenses 308 which are capable of collecting and delivering light to microparticles fixed to substrate 300. Excitation beam 310 from light source 312, which is preferably a laser, is directed to beam splitter 314, e.g. a

5   dichroic mirror, which re-directs the beam through microscope 306 and objective lens 308 which, in turn, focuses the beam onto substrate 300. Lens 308 collects fluorescence 316 emitted from the microparticles and directs it through beam splitter 314 to signal distribution optics 318 which, in turn, directs fluorescence to one or more suitable opto-electronic devices for converting some fluorescence characteristic,

10   e.g. intensity, lifetime, or the like, to an electrical signal. Signal distribution optics 318 may comprise a variety of components standard in the art, such as bandpass filters, fiber optics, rotating mirrors, fixed position mirrors and lenses, diffraction gratings, and the like. As illustrated in Figure 2, signal distribution optics 318 directs fluorescence 316 to four separate photomultiplier tubes, 330, 332, 334, and 336,

15   whose output is then directed to pre-amps and photon counters 350, 352, 354, and 356. The output of the photon counters is collected by computer 304, where it can be stored, analyzed, and viewed on video 360. Alternatively, signal distribution optics 318 could be a diffraction grating which directs fluorescent signal 318 onto a CCD array.

20   The stability and reproducibility of the positional localization in scanning will determine, to a large extent, the resolution for separating closely spaced microparticles. Preferably, the scanning systems should be capable of resolving closely spaced microparticles, e.g. separated by a particle diameter or less. Thus, for most applications, e.g. using CPG microparticles, the scanning system should at least

25   have the capability of resolving objects on the order of 10-100 μm. Even higher resolution may be desirable in some embodiments, but with increase resolution, the time required to fully scan a substrate will increase; thus, in some embodiments a compromise may have to be made between speed and resolution. Increases in scanning time can be achieved by a system which only scans positions where

30   microparticles are known to be located, e.g from an initial full scan. Preferably, microparticle size and scanning system resolution are selected to permit resolution of fluorescently labeled microparticles randomly disposed on a plane at a density between about ten thousand to one hundred thousand microparticles per cm^2.

In sequencing applications, loaded microparticles can be fixed to the surface

35   of a substrate in variety of ways. The fixation should be strong enough to allow the microparticles to undergo successive cycles of reagent exposure and washing without significant loss. When the substrate is glass, its surface may be derivatized with an alkylamino linker using commercially available reagents, e.g. Pierce Chemical, which

in turn may be cross-linked to avidin, again using conventional chemistries, to form an avidinated surface. Biotin moieties can be introduced to the loaded microparticles in a number of ways. For example, a fraction, e.g. 10-15 percent, of the cloning vectors used to attach tags to polynucleotides are engineered to contain a unique

5   restriction site (providing sticky ends on digestion) immediately adjacent to the polynucleotide insert at an end of the polynucleotide opposite of the tag. The site is excised with the polynucleotide and tag for loading onto microparticles. After loading, about 10-15 percent of the loaded polynucleotides will possess the unique restriction site distal from the microparticle surface. After digestion with the

10   associated restriction endonuclease, an appropriate double stranded adaptor containing a biotin moiety is ligated to the sticky end. The resulting microparticles are then spread on the avidinated glass surface where they become fixed via the biotin-avidin linkages.

Alternatively and preferably when sequencing by ligation is employed, in the

15   initial ligation step a mixture of probes is applied to the loaded microparticle: a fraction of the probes contain a type IIs restriction recognition site, as required by the sequencing method, and a fraction of the probes have no such recognition site, but instead contain a biotin moiety at its non-ligating end. Preferably, the mixture - comprises about 10-15 percent of the biotinylated probe.

20   In still another alternative, when DNA-loaded microparticles are applied to a glass substrate, the DNA may nonspecifically adsorb to the glass surface upon several hours, e.g. 24 hours, incubation to create a bond sufficiently strong to permit repeated exposures to reagents and washes without significant loss of microparticles. Preferably, such a glass substrate is a flow cell, which may comprise a channel etched

25   in a glass slide. Preferably, such a channel is closed so that fluids may be pumped through it and has a depth sufficiently close to the diameter of the microparticles so that a monolayer of microparticles is trapped within a defined observation region.

<div align="center">

Identification of Novel Polynucleotides

in cDNA Libraries
</div>

30   Novel polynucleotides in a cDNA library can be identified by constructing a library of cDNA molecules attached to microparticles, as described above. A large fraction of the library, or even the entire library, can then be partially sequenced in parallel. After isolation of mRNA, and perhaps normalization of the population as

35   taught by Soares et al, Proc. Natl. Acad. Sci., 91: 9228-9232 (1994), or like references, the following primer may by hybridized to the polyA tails for first strand synthesis with a reverse transcriptase using conventional protocols (SEQ ID NO: 1):

```
5'-mRNA- [A]n -3'
         [T]19-[primer site]-GG[W,W,W,C]9ACCAGCTGATC-5'
```

where [W,W,W,C]9 represents a tag as described above, "ACCAGCTGATC" is an

5   optional sequence forming a restriction site in double stranded form, and "primer site"
is a sequence common to all members of the library that is later used as a primer
binding site for amplifying polynucleotides of interest by PCR.

       After reverse transcription and second strand synthesis by conventional
techniques, the double stranded fragments are inserted into a cloning vector as

10   described above and amplified. The amplified library is then sampled and the sample
amplified. The cloning vectors from the amplified sample are isolated, and the tagged
cDNA fragments excised and purified. After rendering the tag single stranded with a
polymerase as described above, the fragments are methylated and sorted onto
microparticles in accordance with the invention. Preferably, as described above, the

15   cloning vector is constructed so that the tagged cDNAs can be excised with an
endonuclease, such as Fok I, that will allow immediate sequencing by the preferred
single base method after sorting and ligation to microparticles.

       Stepwise sequencing is then carried out simultaneously on the whole library,
or one or more large fractions of the library, in accordance with the invention until a

20   sufficient number of nucleotides are identified on each cDNA for unique
representation in the genome of the organism from which the library is derived. For
example, if the library is derived from mammalian mRNA then a randomly selected
sequence 14-15 nucleotides long is expected to have unique representation among the
2-3 thousand megabases of the typical mammalian genome. Of course identification

25   of far fewer nucleotides would be sufficient for unique representation in a library
derived from bacteria, or other lower organisms. Preferably, at least 20-30
nucleotides are identified to ensure unique representation and to permit construction
of a suitable primer as described below. The tabulated sequences may then be
compared to known sequences to identify unique cDNAs.

30   Unique cDNAs are then isolated by conventional techniques, e.g. constructing
a probe from the PCR amplicon produced with primers directed to the prime site and
the portion of the cDNA whose sequence was determined. The probe may then be
used to identify the cDNA in a library using a conventional screening protocol.

       The above method for identifying new cDNAs may also be used to fingerprint

35   mRNA populations, either in isolated measurements or in the context of a
dynamically changing population. Partial sequence information is obtained
simultaneously from a large sample, e.g. ten to a hundred thousand, or more, of
cDNAs attached to separate microparticles as described in the above method.

Example 1

Construction of a Tag Library

An exemplary tag library is constructed as follows to form the chemically

5   synthesized 9-word tags of nucleotides A, G, and T defined by the formula:

$$3'\text{-TGGC-}[^4(A,G,T)_9]\text{-CCCCp}$$

where "$[^4(A,G,T)_9]$" indicates a tag mixture where each tag consists of nine 4-mer

10   words of A, G, and T; and "p" indicate a 5' phosphate. This mixture is ligated to the

following right and left primer binding regions (SEQ ID NO: 4 and SEQ ID NO 5):


```
        5'- AGTGGCTGGGCATCGGACCG          5'- GGGGCCCAGTCAGCGTCGAT
            TCACCGACCCGTAGCCp                 GGGTCAGTCGCAGCTA
```

15

                        LEFT   RIGHT

The right and left primer binding regions are ligated to the above tag mixture, after

which the single stranded portion of the ligated structure is filled with DNA

20   polymerase then mixed with the right and left primers indicated below and amplified

to give a tag library (SEQ ID NO: 6).


```
        Left Primer
25
        5'- AGTGGCTGGGCATCGGACCG

        5'- AGTGGCTGGGCATCGGACCG-  [4(A,G,T)9]-GGGGCCCAGTCAGCGTCGAT
            TCACCGACCCGTAGCCTGGC-  [4(A,G,T)9]-CCCCGGGTCAGTCGCAGCTA
30
                                          CCCCGGGTCAGTCGCAGCTA-5'

                                          Right Primer
```


35   The underlined portion of the left primer binding region indicates a Rsr II recognition

site. The left-most underlined region of the right primer binding region indicates

recognition sites for Bsp 120I, Apa I, and Eco O 109I, and a cleavage site for Hga I.

The right-most underlined region of the right primer binding region indicates the

recognition site for Hga I. Optionally, the right or left primers may be synthesized

40   with a biotin attached (using conventional reagents, e.g. available from Clontech

Laboratories, Palo Alto, CA) to facilitate purification after amplification and/or

cleavage.

NOT FURNISHED UPON FILING

```
                      primer binding site              Ppu MI site
                              ↓                             ↓
        (plasmid)-5' -AAAAGGAGGAGGCCTTGATAGAGAGGACCT- ⌐
                     -TTTTCCTCCTCCGGAACTATCTCTCCTGGA- ⌐
```

```
                         primer binding site
                                 ↓
        -GTTTAAAC-GGATCC-TCTTCCTCTTCCTCTTCC-3'-(plasmid)
        -CAAATTTG-CCTAGG-AGAAGGAGAAGGAGAAGG-
             ↑        ↑
                   Bam HI site
           Pme I site
```

The plasmid is cleaved with Ppu MI and Pme I (to give a Rsr II-compatible end and a flush end so that the insert is oriented) and then methylated with DAM methylase. The tag-containing construct is cleaved with Rsr II and then ligated to the open plasmid, after which the conjugate is cleaved with Mbo I and Bam HI to permit ligation and closing of the plasmid. The plasmid is then amplified and isolated and used in accordance with the invention.

Example 3

Changes in Gene Expression Profiles in Liver Tissue of Rats
Exposed to Various Xenobiotic Agents

In this experiment, to test the capability of the method of the invention to detect genes induced as a result of exposure to xenobiotic compounds, the gene expression profile of rat liver tissue is examined following administration of several compounds known to induce the expression of cytochrome P-450 isoenzymes. The results obtained from the method of the invention are compared to results obtained from reverse transcriptase PCR measurements and immunochemical measurements of the cytochrome P-450 isoenzymes. Protocols and materials for the latter assays are described in Morris et al, Biochemical Pharmacology, 52: 781-792 (1996).

Male Sprague-Dawley rats between the ages of 6 and 8 weeks and weighing 200-300 g are used, and food and water are available to the animals *ad lib*. Test compounds are phenobarbital (PB), metyrapone (MET), dexamethasone (DEX), clofibrate (CLO), corn oil (CO), and β-naphthoflavone (BNF), and are available from Sigma Chemical Co. (St. Louis, MO). Antibodies against specific P-450 enzymes are available from the following sources: rabbit anti-rat CYP3A1 from Human Biologics, Inc. (Phoenix, AZ); goat anti-rat CYP4A1 from Daiichi Pure Chemicals Co. (Tokyo,

Japan); monoclonal mouse anti-rat CYP1A1, monoclonal mouse anti-rat CYP2C11, goat anti-rat CYP2E1, and monoclonal mouse anti-rat CYP2B1 from Oxford Biochemical Research, Inc. (Oxford, MI). Secondary antibodies (goat anti-rabbit IgG, rabbit anti-goat IgG and goat anti-mouse IgG) are available from Jackson

5   ImmunoResearch Laboratories (West Grove, PA).

Animals are administered either PB (100 mg/kg), BNF (100 mg/kg), MET (100 mg/kg), DEX (100 mg/kg), or CLO (250 mg/kg) for 4 consecutive days via intraperitoneal injection following a dosing regimen similar to that described by Wang et al, Arch. Biochem. Biophys. 290: 355-361 (1991). Animals treated with

10   H_2O and CO are used as controls. Two hours following the last injection (day 4), animals are killed, and the livers are removed. Livers are immediately frozen and stored at -70°C.

Total RNA is prepared from frozen liver tissue using a modification of the method described by Xie et al, Biotechniques, 11: 326-327 (1991). Approximately

15   100-200 mg of liver tissue is homogenized in the RNA extraction buffer described by Xie et al to isolate total RNA. The resulting RNA is reconstituted in diethylpyrocarbonate-treated water, quantified spectrophotometrically at 260 nm, and adjusted to a concentration of 100 μg/ml. Total RNA is stored in - diethylpyrocarbonate-treated water for up to 1 year at -70°C without any apparent

20   degradation. RT-PCR and sequencing are performed on samples from these preparations.

For sequencing, samples of RNA corresponding to about 0.5 μg of poly(A)$^+$ RNA are used to construct libraries of tag-cDNA conjugates following the protocol described in the section entitled "Attaching Tags to Polynucleotides for Sorting onto

25   Solid Phase Supports," with the following exception: the tag repertoire is constructed from six 4-nucleotide words from Table II. Thus, the complexity of the repertoire is 8^6 or about 2.6×10^5. For each tag-cDNA conjugate library constructed, ten samples of about ten thousand clones are taken for amplification and sorting. Each of the amplified samples is separately applied to a fixed monolayer of about 10^6 10 μm

30   diameter GMA beads containing tag complements. That is, the "sample" of tag complements in the GMA bead population on each monolayer is about four fold the total size of the repertoire, thus ensuring there is a high probability that each of the sampled tag-cDNA conjugates will find its tag complement on the monolayer. After the oligonucleotide tags of the amplified samples are rendered single stranded as

35   described above, the tag-cDNA conjugates of the samples are separately applied to the monolayers under conditions that permit specific hybridization only between oligonucleotide tags and tag complements forming perfectly matched duplexes. Concentrations of the amplified samples and hybridization times are selected to

permit the loading of about 5×10^4 to 2×10^5 tag-cDNA conjugates on each bead where perfect matches occur. After ligation, 9-12 nucleotide portions of the attached cDNAs are determined in parallel by the single base sequencing technique described by Brenner in International patent application PCT/US95/03678. Frequency

5   distributions for the gene expression profiles are assembled from the sequence information obtained from each of the ten samples.

RT-PCRs of selected mRNAs corresponding to cytochrome P-450 genes and the constitutively expressed cyclophilin gene are carried out as described in Morris et al (cited above). Briefly, a 20 μL reaction mixture is prepared containing 1x reverse

10   transcriptase buffer (Gibco BRL), 10 nM dithiothreitol, 0.5 nM dNTPs, 2.5 μM oligo $d(T)_{15}$ primer, 40 units RNasin (Promega, Madison, WI), 200 units RNase H-reverse transcriptase (Gibco BRL), and 400 ng of total RNA (in diethylpyrocarbonate-treated water). The reaction is incubated for 1 hour at 37°C followed by inactivation of the enzyme at 95°C for 5 min. The resulting cDNA is stored at -20°C until used. For

15   PCR amplification of cDNA, a 10 μL reaction mixture is prepared containing 10x polymerase reaction buffer, 2 mM $MgCl_2$, 1 unit Taq DNA polymerase (Perkin-Elmer, Norwalk, CT), 20 ng cDNA, and 200 nM concentration of the 5' and 3' specific PCR primers of the sequences described in Morris et al (cited above). PCRs -are carried out in a Perkin-Elmer 9600 thermal cycler for 23 cycles using melting,

20   annealing, and extension conditions of 94°C for 30 sec., 56°C for 1 min., and 72°C for 1 min., respectively. Amplified cDNA products are separated by PAGE using 5% native gels. Bands are detected by staining with ethidium bromide.

Western blots of the liver proteins are carried out using standard protocols after separation by SDS-PAGE. Briefly, proteins are separated on 10% SDS-PAGE

25   gels under reducing conditions and immunoblotted for detection of P-450 isoenzymes using a modification of the methods described in Harris et al, Proc. Natl. Acad. Sci., 88: 1407-1410 (1991). Protein are loaded at 50 μg/lane and resolved under constant current (250 V) for approximately 4 hours at 2°C. Proteins are transferred to nitrocellulose membranes (Bio-Rad, Hercules, CA) in 15 mM Tris buffer containing

30   120 mM glycine and 20% (v/v) methanol. The nitrocellulose membranes are blocked with 2.5% BSA and immunoblotted for P-450 isoenzymes using primary monoclonal and polyclonal antibodies and secondary alkaline phosphatase conjugated anti-IgG. Immunoblots are developed with the Bio-Rad alkaline phosphatase substrate kit.

The three types of measurements of P-450 isoenzyme induction showed

35   substantial agreement.

APPENDIX Ia

Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```
Program minxh
c
c
c
        integer*2 sub1(6),mset1(1000,6),mset2(1000,6)
        dimension nbase(6)
c
c
        write(*,*)'ENTER SUBUNIT LENGTH'
        read(*,100)nsub
100     format(i1)
        open(1,file='sub4.dat',form='formatted',status='new')
c
c
      nset=0
      do 7000 m1=1,3
         do 7000 m2=1,3
            do 7000 m3=1,3
               do 7000 m4=1,3
                  sub1(1)=m1
                  sub1(2)=m2
                  sub1(3)=m3
                  sub1(4)=m4
c
c
      ndiff=3
c
c
c               Generate set of subunits differing from
c               sub1 by at least ndiff nucleotides.
.c              Save in mset1.
c
c
      jj=1
      do 900 j=1,nsub
900      mset1(1,j)=sub1(j)
c
c
      do 1000 k1=1,3
         do 1000 k2=1,3
            do 1000 k3=1,3
               do 1000 k4=1,3
c
c
                  nbase(1)=k1
                  nbase(2)=k2
                  nbase(3)=k3
                  nbase(4)=k4
```

```
c
               n=0
               do 1200 j=1,nsub
                  if(sub1(j).eq.1 .and. nbase(j).ne.1 .or.
      1              sub1(j).eq.2 .and. nbase(j).ne.2 .or.
      3              sub1(j).eq.3 .and. nbase(j).ne.3) then
                  n=n+1
                  endif
1200           continue
c
c
          if(n.ge.ndiff) then
c
c
c                                        If number of mismatches
c                                        is greater than or equal
c                                        to ndiff then record
c                                        subunit in matrix mset
c
c
               jj=jj+1
                do 1100 i=1,nsub
1100               mset1(jj,i)=nbase(i)
               endif
c
c
1000      continue
c
c
               do 1325 j2=1,nsub
               mset2(1,j2)=mset1(1,j2)
1325           mset2(2,j2)=mset1(2,j2)
c
c
c                                        Compare subunit 2 from
c                                        mset1 with each successive
c                                        subunit in mset1, i.e. 3,
c                                        4,5, ... etc. Save those
c                                        with mismatches .ge. ndiff
c                                        in matrix mset2 starting at
c                                        position 2.
c                                          Next transfer contents
c                                        of mset2 into mset1 and
c                                        start
c                                        comparisons again this time
c                                        starting with subunit 3.
c                                        Continue until all subunits
c                                        undergo the comparisons.
c
c
          npass=0
c
c
1700      continue
          kk=npass+2
          npass=npass+1
c
```

```
c
                do 1500 m=npass+2,jj
                   n=0
                   do 1600 j=1,nsub
                      if(mset1(npass+1,j).eq.1.and.mset1(m,j).ne.1.or.
       2                 mset1(npass+1,j).eq.2.and.mset1(m,j).ne.2.or.
       2                 mset1(npass+1,j).eq.3.and.mset1(m,j).ne.3) then
                         n=n+1
                      endif
1600                  continue
                   if(n.ge.ndiff) then
                         kk=kk+1
                         do 1625 i=1,nsub
1625                        mset2(kk,i)=mset1(m,i)
                   endif
1500               continue
c
c                                              kk is the number of subunits
c                                              stored in mset2
c
c                                              Transfer contents of mset2
c                                              into mset1 for next pass.
c
c
                 do 2000 k=1,kk
                    do 2000 m=1,nsub
2000                   mset1(k,m)=mset2(k,m)
                 if(kk.lt.jj) then
                    jj=kk
                    goto 1700
                    endif
c
c
                nset=nset+1
                write(1,7009)
7009             format(/)
                do 7008 k=1,kk
7008                write(1,7010)(mset1(k,m),m=1,nsub)
7010            format(4i1)
                write(*,*)
                write(*,120) kk,nset
120             format(1x,'Subunits in set=',i5,2x,'Set No=',i5)
7000             continue
                 close(1)
c
c
                end
c               ******************************
c               ******************************
```

APPENDIX Ib

Exemplary computer program for generating
minimally cross hybridizing sets
(single stranded tag/single stranded tag complement)

```
Program tagN
c
c
c           Program tagN generates minimally cross-hybridizing
c              sets of subunits given i) N--subunit length, and ii)
c              an initial subunit sequence.  tagN assumes that only
c              3 of the four natural nucleotides are used in the tags.
c
c

            character*1 sub1(20)
            integer*2 mset(10000,20), nbase(20)
c
c

            write(*,*)'ENTER SUBUNIT LENGTH'
            read(*,100)nsub
100         format(i2)
c
c

            write(*,*)'ENTER SUBUNIT SEQUENCE'
            read(*,110)(sub1(k),k=1,nsub)
110         format(20a1)
c
c

            ndiff=10
c
c
c                       Let a=1 c=2 g=3 & t=4
c
c

         do 800 kk=1,nsub
         if(sub1(kk).eq.'a') then
            mset(1,kk)=1
            endif
               if(sub1(kk).eq.'c') then
                  mset(1,kk)=2
                  endif
                     if(sub1(kk).eq.'g') then
                        mset(1,kk)=3
                        endif
                           if(sub1(kk).eq.'t') then
                              mset(1,kk)=4
                              endif
800      continue
c
c
c                    Generate set of subunits differing from
c                    sub1 by at least ndiff nucleotides.
c
c

         jj=1
c
c

         do 1000 kl=1,3
```

```
                 do 1000 k2=1,3
                    do 1000 k3=1,3
                       do 1000 k4=1,3
                          do 1000 k5=1,3
                             do 1000 k6=1,3
                                do 1000 k7=1,3
                                   do 1000 k8=1,3
                                      do 1000 k9=1,3
                                         do 1000 k10=1,3
                 do 1000 k11=1,3
                    do 1000 k12=1,3
                       do 1000 k13=1,3
                          do 1000 k14=1,3
                             do 1000 k15=1,3
                                do 1000 k16=1,3
                                   do 1000 k17=1,3
                                      do 1000 k18=1,3
                                         do 1000 k19=1,3
                                            do 1000 k20=1,3
c
c

                    nbase(1)=k1
                    nbase(2)=k2
                    nbase(3)=k3
                    nbase(4)=k4
                    nbase(5)=k5
                    nbase(6)=k6
                    nbase(7)=k7
                    nbase(8)=k8
                    nbase(9)=k9
                    nbase(10)=k10
                    nbase(11)=k11
                    nbase(12)=k12
                    nbase(13)=k13
                    nbase(14)=k14
                    nbase(15)=k15
                    nbase(16)=k16
                    nbase(17)=k17
                    nbase(18)=k18
                    nbase(19)=k19
                    nbase(20)=k20
c
c

                 do 1250 nn=1,jj
c
                  n=0
                  do 1200 j=1,nsub
                    if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                 mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
     2                 mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3                 mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                      n=n+1
                    endif
1200              continue
c
c
           if(n.lt.ndiff) then
              goto 1000
              endif
1250       continue
c
c

                 jj=jj+1
                 write(*,130)(nbase(i),i=1,nsub),jj
                  do 1100 i=1,nsub
```

```
                         mset(jj,i)=nbase(i)
      1100                   continue
      c
      c
      1000      continue
      c
      c
                write(*,*)
      130       format(10x,20(1x,i1),5x,i5)
                write(*,*)
                write(*,120) jj
      120       format(1x,'Number of words=',i5)
      c
      c
                end
      c
      c            *****************************************
      c            *****************************************
```

APPENDIX Ic

Exemplary computer program for generating

minimally cross hybridizing sets

(double stranded tag/single stranded tag complement)

```
Program 3tagN
c
c
c           Program 3tagN generates minimally cross-hybridizing
c               sets of duplex subunits given i) N--subunit length,
c               and ii) an initial homopurine sequence.
c
c
            character*1 sub1(20)
            integer*2 mset(10000,20), nbase(20)
c
c
            write(*,*)'ENTER SUBUNIT LENGTH'
            read(*,100)nsub
·100        format(i2)
c
c
            write(*,*)'ENTER SUBUNIT SEQUENCE a & g only'
            read(*,110)(sub1(k),k=1,nsub)
110         format(20a1)
c
·           ndiff=10
c
c                   Let a=1 and g=2
c
         do 800 kk=1,nsub
         if(sub1(kk).eq."a') then
            mset(1,kk)=1
            endif
                if(sub1(kk).eq.'g') then
                    mset(1,kk)=2
                    endif
800      continue
c

         jj=1
c
         do 1000 k1=1,3
           do 1000 k2=1,3
             do 1000 k3=1,3
               do 1000 k4=1,3
                 do 1000 k5=1,3
                   do 1000 k6=1,3
                     do 1000 k7=1,3
                       do 1000 k8=1,3
                         do 1000 k9=1,3
                           do 1000 k10=1,3
         do 1000 k11=1,3
           do 1000 k12=1,3
             do 1000 k13=1,3
               do 1000 k14=1,3
                 do 1000 k15=1,3
                   do 1000 k16=1,3
                     do 1000 k17=1,3
                       do 1000 k18=1,3
```

```
                              do 1000 k19=1,3
                                do 1000 k20=1,3
c
                        nbase(1)=k1
                        nbase(2)=k2
                        nbase(3)=k3
                        nbase(4)=k4
                        nbase(5)=k5
                        nbase(6)=k6
                        nbase(7)=k7
                        nbase(8)=k8
                        nbase(9)=k9
                        nbase(10)=k10
                        nbase(11)=k11
                        nbase(12)=k12
                        nbase(13)=k13
                        nbase(14)=k14
                        nbase(15)=k15
                        nbase(16)=k16
                        nbase(17)=k17
                        nbase(18)=k18
                        nbase(19)=k19
                        nbase(20)=k20
c
                  do 1250 nn=1,jj
c
                   n=0
                   do 1200 j=1,nsub
                     if(mset(nn,j).eq.1 .and. nbase(j).ne.1 .or.
     1                  mset(nn,j).eq.2 .and. nbase(j).ne.2 .or.
     2                  mset(nn,j).eq.3 .and. nbase(j).ne.3 .or.
     3                  mset(nn,j).eq.4 .and. nbase(j).ne.4) then
                       n=n+1
                     endif
1200                 continue
c
            if(n.lt.ndiff) then
               goto 1000
               endif
1250        continue
c
                  jj=jj+1
                  write(*,130)(nbase(i),i=1,nsub),jj
                   do 1100 i=1,nsub
                      mset(jj,i)=nbase(i)
1100                  continue
c
1000        continue
c
            write(*,*)
130         format(10x,20(1x,i1),5x,i5)
            write(*,*)
            write(*,120) jj
120         format(1x,'Number of words=',i5)
c
c
            end
```

SEQUENCE LISTING

(1) GENERAL INFORMATION:

(i) APPLICANT: David W. Martin, Jr.

(ii) TITLE OF INVENTION: Measurement of Gene Expression profiles in
Toxicity Determination

(iii) NUMBER OF SEQUENCES: 7

(iv) CORRESPONDENCE ADDRESS:
     (A) ADDRESSEE: Stephen C. Macevicz, Lynx Therapeutics, Inc.
     (B) STREET: 3832 Bay Center Place
     (C) CITY: Hayward
     (D) STATE: California
     (E) COUNTRY: USA
     (F) ZIP: 94545

(v) COMPUTER READABLE FORM:
     (A) MEDIUM TYPE: 3.5 inch diskette
     (B) COMPUTER: IBM compatible
     (C) OPERATING SYSTEM: Windows 3.1
     (D) SOFTWARE: Microsoft Word 5.1

(vi) CURRENT APPLICATION DATA:
     (A) APPLICATION NUMBER:
     (B) FILING DATE:
     (C) CLASSIFICATION:

(vii) PRIOR APPLICATION DATA:
     (A) APPLICATION NUMBER: PCT/US96/09513
     (B) FILING DATE: 06-JUN-96

(vii) PRIOR APPLICATION DATA:
     (A) APPLICATION NUMBER: PCT/US95/12791
     (B) FILING DATE: 12-OCT-95

(viii) ATTORNEY/AGENT INFORMATION:
     (A) NAME: Stephen C. Macevicz
     (B) REGISTRATION NUMBER: 30,285
     (C) REFERENCE/DOCKET NUMBER: 813wo

(ix) TELECOMMUNICATION INFORMATION:
     (A) TELEPHONE: (510) 670-9365
     (B) TELEFAX: (510) 670-9302

(2) INFORMATION FOR SEQ ID NO: 1:

(i) SEQUENCE CHARACTERISTICS:
     (A) LENGTH: 11 nucleotides
     (B) TYPE: nucleic acid
     (C) STRANDEDNESS: single
     (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 1:

CTAGTCGACC A   11

(2) INFORMATION FOR SEQ ID NO: 2:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 11 nucleotides
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 2:

NRRGATCYNN N   11

(2) INFORMATION FOR SEQ ID NO: 3:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 38 nucleotides
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: single
      (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 3:

GAGGATGCCT TTATGGATCC ACTCGAGATC CCAATCCA   38

(2) INFORMATION FOR SEQ ID NO: 4:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 20 nucleotides
      (B) TYPE: nucleic acid
      (C) STRANDEDNESS: double
      (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 4:

AGTGGCTGGG CATCGGACCG   20

(2) INFORMATION FOR SEQ ID NO: 5:

(i) SEQUENCE CHARACTERISTICS:
      (A) LENGTH: 20 nucleotides
      (B) TYPE: nucleic acid

          (C) STRANDEDNESS: double
          (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 5:

GGGGCCCAGT CAGCGTCGAT   20

(2) INFORMATION FOR SEQ ID NO: 6:

(i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 20 nucleotides
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: single
          (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 6:

ATCGACGCTG ACTGGGCCCC   16

(2) INFORMATION FOR SEQ ID NO: 7:

(i) SEQUENCE CHARACTERISTICS:
          (A) LENGTH: 62 nucleotides
          (B) TYPE: nucleic acid
          (C) STRANDEDNESS: double
          (D) TOPOLOGY: linear

(xi) SEQUENCE DESCRIPTION: SEQ ID NO: 7:

AAAAGGAGGA GGCCTTGATA GAGAGGACCT GTTTAAACGG ATCCTCTTCC   50
TCTTCCTCTT CC   62

I claim:

1.   A method of determining the toxicity of a compound, the method comprising the steps of:

5   administering the compound to a test organism;

        extracting a population of mRNA molecules from each of one or more tissues of the test organism;

        forming a separate population of cDNA molecules from each population of mRNA molecules from the one or more tissues such that each cDNA molecule of a

10   separate population has an oligonucleotide tag attached, the oligonucleotide tags being selected from the same minimally cross-hybridizing set;

        separately sampling each population of cDNA molecules such that substantially all different cDNA molecules within a separate population have different oligonucleotide tags attached;

15   sorting the cDNA molecules of each separate population by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

        determining the nucleotide sequence of a portion of each of the sorted cDNA

20   molecules of each separate population to form a frequency distribution of expressed genes for each of the one or more tissues; and

        correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

25 2.   The method of claim 1 wherein said oligonucleotide tag and said complement of said oligonucleotide tag are single stranded.

3.   The method of claim 2 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in

30   length and each subunit being selected from the same minimally cross-hybridizing set.

4.   The method of claim 3 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation

35   of unloaded microparticles.

5.   The method of claim 4 further including a step of separating said loaded microparticles from said unloaded microparticles.

6.   The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

5

7.   The method of claim 6 wherein said number of loaded microparticles is at least 100,000.

8.   The method of claim 7 wherein said number of loaded microparticles is at
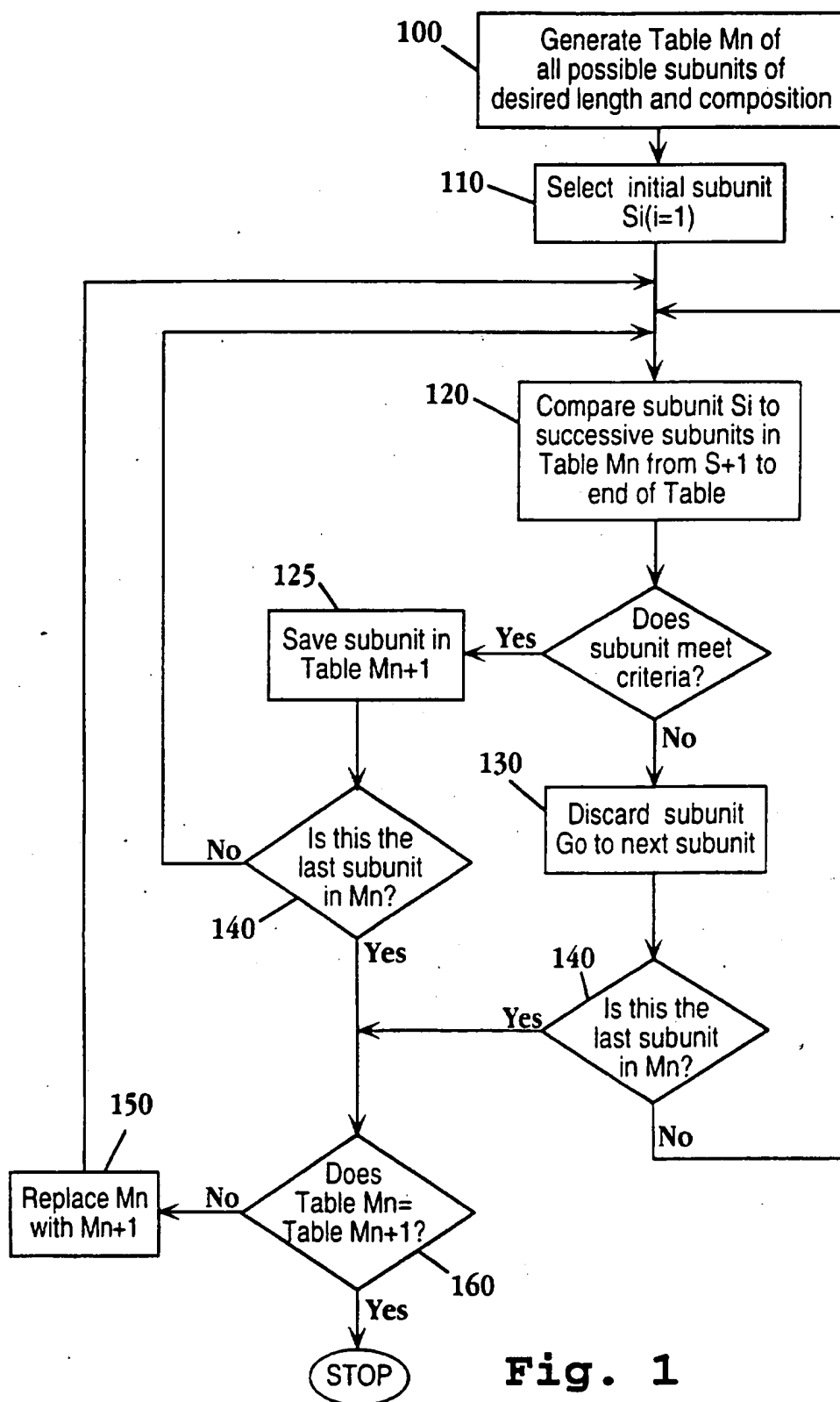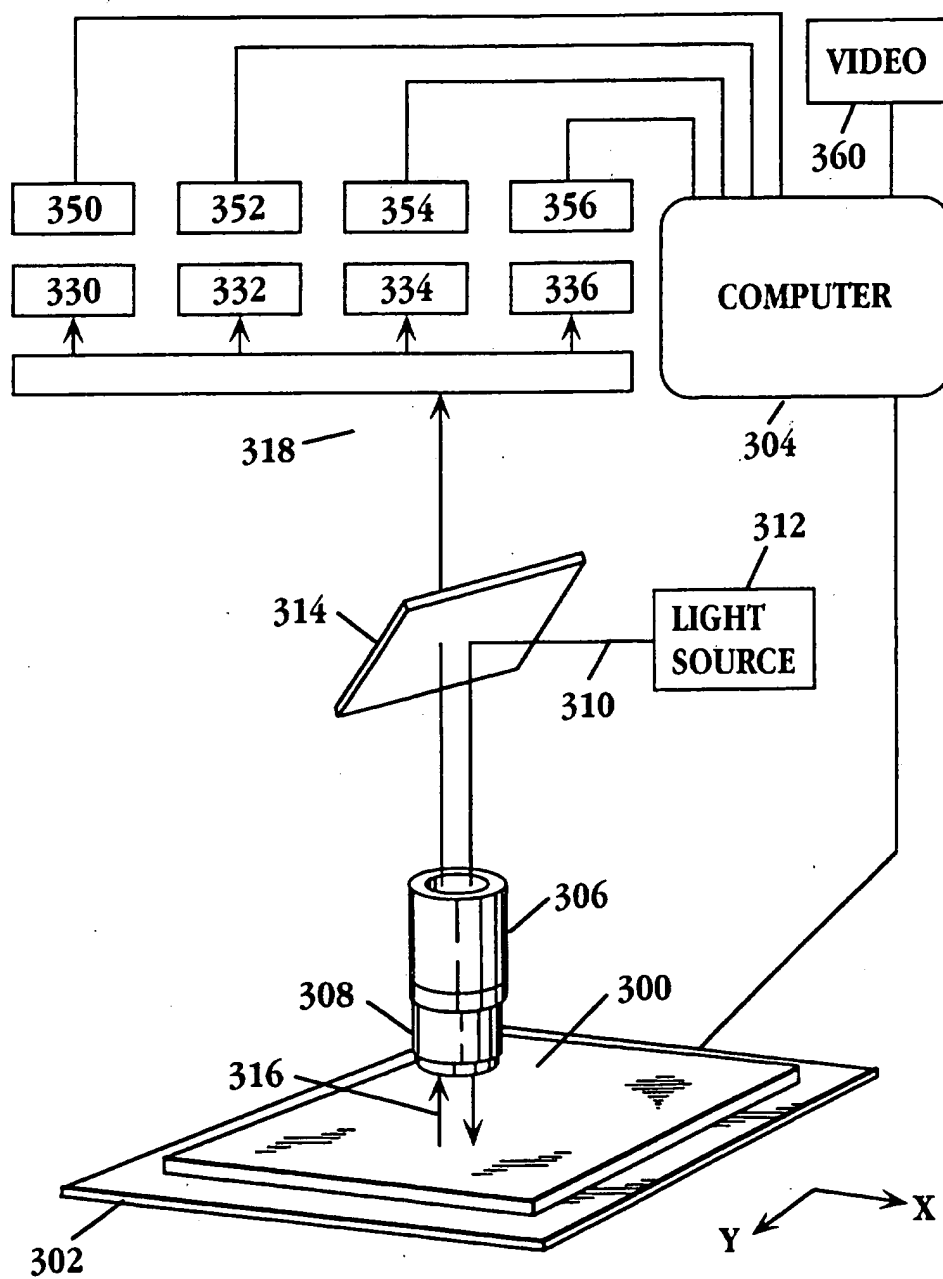10   least 500,000.

9.   The method of claim 5 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule
15   present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

10.   The method of claim 4 wherein said test organism is a mammalian tissue
- culture.
20

11.   The method of claim 10 wherein said mammalian tissue culture comprises hepatocytes.

12.   The method of claim 4 wherein said test organism is an animal selected from
25   the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

13.   The method of claim 12 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large
30   intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

14.   A method of identifying genes which are differentially expressed in a selected
35   tissue of a test animal after treatment with a compound, the method comprising the steps of:
        administering the compound to a test animal;

extracting a population of mRNA molecules from the selected tissue of the test animal;

forming a population of cDNA molecules from the population of mRNA molecules such that each cDNA molecule has an oligonucleotide tag attached, the

5   oligonucleotide tags being selected from the same minimally cross-hybridizing set;

sampling the population of cDNA molecules such that substantially all different cDNA molecules have different oligonucleotide tags attached;

sorting the cDNA molecules by specifically hybridizing the oligonucleotide tags with their respective complements, the respective complements being attached as

10   uniform populations of substantially identical complements in spatially discrete regions on one or more solid phase supports;

determining the nucleotide sequence of a portion of each of the sorted cDNA molecules to form a frequency distribution of expressed genes; and

identifying genes expressed in response to administering the compound by

15   comparing the frequencing distribution of expressed genes of the selected tissue of the test animal with a frequency distribution of expressed genes of the selected tissue of a control animal.

15.   The method of claim 14 wherein said oligonucleotide tag and said

20   complement of said oligonucleotide tag are single stranded.

16.   The method of claim 15 wherein said oligonucleotide tag consists of a plurality of subunits, each subunit consisting of an oligonucleotide of 3 to 9 nucleotides in length and each subunit being selected from the same minimally cross-

25   hybridizing set.

17.   The method of claim 16 wherein said one or more solid phase supports are microparticles and wherein said step of sorting said cDNA molecules onto the microparticles produces a subpopulation of loaded microparticles and a subpopulation

30   of unloaded microparticles.

18.   The method of claim 17 further including a step of separating said loaded microparticles from said unloaded microparticles.

35   19.   The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is at least 10,000.

20.   The method of claim 19 wherein said number of loaded microparticles is at least 100,000.

21.   The method of claim 20 wherein said number of loaded microparticles is at
5   least 500,000.

22.   The method of claim 18 further including a step of repeating said steps of sampling, sorting, and separating until a number of said loaded microparticles is accumulated is sufficient to estimate the relative abundance of a cDNA molecule
10   present in said population at a frequency within the range of from 0.1% to 5% with a 95% confidence limit no larger than 0.1% of said population.

23.   The method of claim 17 wherein said test animal is selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.
15

24.   The method of claim 23 wherein said selected tissue is selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and
- mesenteric lymph nodes.
20

25.   A use of the technique of massively parallel signature sequencing to determine the toxicity of a compound in a test organism, the use comprising the steps of:
       administering the compound to a test organism;
       extracting a population of mRNA molecules from each of one or more tissues
25   of the test organism and forming a population of cDNA molecules for each of the one or more tissues;
       . determining the nucleotide sequence of a portion of each of the cDNA molecules of each separate population using massively parallel signature sequencing to form a frequency distribution of expressed genes for each of the one or more
30   tissues; and
       correlating the frequency distribution of expressed genes in each of the one or more tissues with the toxicity of the compound.

26.   The use of claim 25 wherein said test organism is a mammalian tissue culture.
35

27.   The use of claim 26 wherein said mammalian tissue culture comprises hepatocytes.

28.   The use of claim 25 wherein said test organism is an animal selected from the group consisting of rats, mice, hamsters, guinea pigs, rabbits, cats, dogs, pigs, and monkeys.

5   29.   The use of claim 28 wherein said one or more tissues are selected from the group consisting of liver, kidney, brain, cardiovascular, thyroid, spleen, adrenal, large intestine, small intestine, pancrease urinary bladder, stomach, ovary, testes, and mesenteric lymph nodes.

10   30.   A use of the technique of massively parallel signature sequencing to identify genes which are differentially expressed in a test organism after treatment with a compound and which are correlated with toxicity of the compound, the use comprising the steps of:

       administering the compound to the test organism;

15   extracting a population of mRNA molecules from a selected tissue of the test organism and forming a population of cDNA molecules;

       determining the nucleotide sequence of a portion of each of the cDNA molecules using massively parallel signature sequencing to form a frequency - distribution of expressed genes;

20   identifying genes expressed in response to administering the compound by comparing the frequencing distribution of expressed genes of the selected tissue of the test organism with a frequency distribution of expressed genes of the selected tissue of a control organism; and

       determining whether the genes expressed in response to administering the 25   compound are correlated with toxicity of the compound in the test organism.

100 — Generate Table Mn of
all possible subunits of
desired length and composition

110 — Select initial subunit
Si(i=1)

120 — Compare subunit Si to
successive subunits in
Table Mn from S+1 to
end of Table

Does
subunit meet
criteria?

Yes → 125 Save subunit in
Table Mn+1

No → 130 Discard subunit
Go to next subunit

Is this the
last subunit
in Mn? 140

No ←

Yes ↓

Is this the
last subunit
in Mn? 140

Yes ←

No

150 Replace Mn
with Mn+1 ← **No** — Does
Table Mn=
Table Mn+1? 160

Yes

STOP

Fig. 1

2/2

Fig. 2

INTERNATIONAL SEARCH REPORT

International application No.

PCT/US96/16342

A. CLASSIFICATION OF SUBJECT MATTER

IPC(6) : C12Q 1/68; C07H 21/04
US CL : 435/6; 536/24.3

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

U.S. : 435/6; 536/24.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

APS, MEDLINE, BIOSIS, CAPLUS, SCISEARCH
search terms: Martin, David W., toxic?, differential?, express?, cDNA, mRNA, RNA, gene#, hybrid?,

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CHETVERIN et al. Oligonucleotide arrays: New concepts and possibilities. Bio/Technology. 12 November 1994, Vol. 12, pages 1093-1099, especially pages 1095-1096.	1-30
A	BRENNER et al. Encoded combinatorial chemistry. Proceedings of the National Academy of Sciences USA. June 1992, Vol. 89, pages 5381-5383.	1-30
A	MATSUBARA et al. cDNA analyses in the human genome project. Gene. 15 December 1993, Vol. 135, No. 1-2, pages 265-274.	1-30

☒ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

*	Special categories of cited documents:	"T"	later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"A"	document defining the general state of the art which is not considered to be of particular relevance		
"E"	earlier document published on or after the international filing date	"X"	document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"L"	document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)	"Y"	document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"O"	document referring to an oral disclosure, use, exhibition or other means		
"P"	document published prior to the international filing date but later than the priority date claimed	"&"	document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
27 JANUARY 1997	**19 FEB 1997**

Name and mailing address of the ISA/US	Authorized officer
Commissioner of Patents and Trademarks Box PCT Washington, D.C. 20231	SCOTT D. PRIEBE
Facsimile No. (703) 305-3230	Telephone No. (703) 308-0196

Form PCT/ISA/210 (second sheet)(July 1992)*

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 95/21944 A1 (SMITHKLINE BEECHAM CORPORATION) 17 August 1995, page 4, lines 1-4, page 5, lines 31-37, page 17, lines 15-27, page 18, lines 30-35, page 20, line 23 to page 21, line 4.	1-30

FOCUS - 17 of 19 DOCUMENTS

August 11, 1997, Monday

SECTION: Financial News

DISTRIBUTION: TO BUSINESS AND MEDICAL EDITORS

LENGTH: 478 words

HEADLINE: Eli Lilly & Co. and Acacia Biosciences Enter Into Research Collaboration;
First Corporate Agreement for Acacia's Genome Reporter Matrix(TM)

DATELINE: RICHMOND, Calif., Aug. 11

BODY:

Acacia Biosciences and Eli Lilly and Company (Lilly) announced today the signing of a joint research collaboration to utilize Acacia's Genome Reporter Matrix(TM) (GRM) to aid in the selection and optimization of lead compounds. Under the collaboration, Acacia will provide chemical and biological profiles on a class of Lilly's compounds for an undisclosed fee.

Acacia's GRM is an assay-based computer modeling system that uses yeast as a miniature ecosystem. The GRM can profile the extent, nature and quantity of any changes in gene expression. Because of the similarities between the yeast and human genome, the system serves as an excellent surrogate for the human body, mimicking the effects induced by a biologically active molecule.

"Using yeast as a model organism for lead optimization makes a lot of sense given the high degree of homology with human metabolic pathways," said William Current of Lilly Research Laboratories. "Acacia's innovative GRM has the potential to provide enormous insight into the therapeutic impact of our compounds and make the drug discovery process more rational. It should substantially accelerate the development process."

"This first agreement with a major pharmaceutical company is an important milestone in the development of Acacia," said Bruce Cohen, President and CEO of Acacia. "The deal is in line with our strategy of establishing alliances that will allow our collaborators to use genomic profiles to identify and optimize compounds within their existing portfolios. In the long run, this technology can be used to characterize large scale combinatorial libraries, predict side effects prior to clinical trials and resurrect drugs that have failed during clinical trials."

The GRM incorporates two critical elements: chemical response profiles and genetic response profiles. The chemical response profiles measure the change in gene expression caused by potential therapeutics and then rank genes with altered expressions by degree of response. The genetic response profiles measure changes in gene expression caused by mutations in the genes encoding potential targets of pharmaceuticals; these genetic response profiles represent gold standards in drug discovery by defining the response profile expected for drugs with perfect selectivity and specificity. By comparing the two profiles, one can analyze a potential drug candidate's ability to mimic the action of a 'perfect' drug.

Acacia Biosciences is a functional genomics company developing proprietary technologies to enhance the speed and efficacy of drug discovery and development. Acacia's Genome Reporter Matrix capitalizes on the latest advances in genomics and combinatorial chemistry to generate comprehensive profiles of drug candidates' in vivo activity.
SOURCE Acacia Biosciences
CONTACT: Bruce Cohen, President and CEO of Acacia Biosciences, 510-669-2330 ext. 103 or Media: Linda Seaton of Feinstein

LOAD-DATE: August 12, 1997

The Bioreactor Market:
Steady Growth Expected

The worldwide market
for all bioreactors was valued
at $175 million for 1997,
and is expected to be
worth $380 million by 2002.

GENETIC ENGINEERING NEWS

BIOTECHNOLOGY · BIOPROCESS · BIORESEARCH · TECHNOLOGY TRANSFER

Pharmagene Raises More Capital for Research on Human Tissues

By Sophia Fox

Pharmagene, the Royston, U.K.-based biopharmaceutical company specialising in the use of human biomaterials for drug discovery research, has raised a further £5 million from a group of investors led by 3i and Abacus Nominees. The funding will enable the company to expand both its human biomaterials collection and its capabilities across a range of proprietary platform technologies.

Gordon Baxter, Ph.D., Pharmagene's cofounder and chief operating officer, claimed, "by the end of this year Pharmagene will have access to the largest collection of human RNAs and proteins anywhere in the world, and a range of innovative, yet robust technologies

Perkin-Elmer Acquires PerSeptive to Expand Its Capabilities in Gene-Based Drug Discovery

By John Sterling

Perkin-Elmer's (PE; Norwalk, CT) decision last month to acquire PerSeptive Biosystems (Framingham, MA) via a $360 million stock swap was designed to strengthen PE in terms of broad capabilities in gene-based drug discovery. The company's main goal is to develop new products to improve the integration of genetic and protein research.

"This merger will enhance our position as an effective provider of innovative, integrated platforms enabling our customers to be more efficient and cost-effective in bringing new pharmaceuticals to market," says Tony L. White, PE's chairman, president and CEO. "The combination of our two companies should bolster our presence in the life sciences, [and it is our] belief that we must take bold action now to lead the emerging era of molecular medicine with leading positions in both genetic and protein analysis."

A driving force behind the merger is the vast amount of genet-

ic information about human disease that is being accumulated by researchers and biotech companies working in the area of genomics. It is becoming increasingly obvious that these data need to be complemented with technologies for

Perkin-Elmer acquired PerSeptive Biosystems for $360 million to obtain new technologies in mass spectrometry, bioseparations and purification for product development projects, spanning the range from genomics to proteomics.

studying proteins and protein networks—a field known as proteomics (see GEN, September 1, 1997, p.1).

PE officials, who claim that MALDI-TOF (Matrix Assisted

FDA OKs Genzyme's Carticel Product for Damage to Knees

Carticel, which was approved for the repair of clinically significant, symptomatic cartilaginous defects of the femoral condyle (medial, lateral or trochlear) caused by acute or repetitive trauma, employs a proprietary process to grow autologous cartilage cells for implantation.

By Naomi Pfeiffer

The FDA has approved a knee-cartilage replacement product made by Genzyme Tissue Repair (Cambridge, MA), a tracking-stock division of Genzyme Corp., for people with trauma-damaged knees.

Carticel™ (autologous cultured chondrocytes) is the first product to be licensed under the FDA's pro-

Strategies for Target Validation Streamline Evaluation of Leads

By Vicki Glaser

Acacia Biosciences (Richmond, CA) last month announced its first agreement with a major pharmaceutical company, signing a deal with Eli Lilly (Indianapolis, IN) to use Acacia's Genome Reporter Matrix (GRM) to select and optimize some of Lilly's lead compounds. Acacia's yeast-based system for profiling drug activity is useful for evaluating the therapeutic potential of lead compounds, and it also has a role in the identification and validation of new drug targets.

"We're using the ecosystem of a cell to allow us to deduce the mechanism of action and target for any chemical," explains Bruce Cohen, president and CEO. "We screen for every target in a cell simultaneously...using transcription as a readout

for how a cell is adapting to any perturbation," he says.

The GRM technology consists of two main databases: one is the genetic response profile, showing the effects of mutations in each individual yeast gene and compensatory gene regulatory mechanisms; the other is the chemical response profile, which documents changes in gene expression in response to chemical compounds. Computational analysis and pattern matching between the genetic and chemical profiles yields information on the specificity, potency and side-effects risk of a drug lead.

Targeting Targets

No longer is mapping and sequencing a gene—or the human genome—an end unto itself, but

Sticky Ends

Avigen received two grants from the **NIH** & **University of California** for research on gene therapy for treatment of cancer & **HIV** infections...**MRL Pharmaceutical Services**, of Reston, **VA**, launched the **TSN** Bug Finder, which is able to locate & retrieve client-specified microorganisms in real-time...**Gensia Sicor, Inc.** will move its corporate staff from San Diego to Irvine, **CA**, by end of year...

FDA accepted **NDA** from **Sepracor** for levalbuterol **HCl** inhalation solution...An $11.7M mezzanine financing has been closed by **Activated Cell Therapy**, which changed its name to **Dendreon Corporation**...**Astra AB** will build major research facility in **Waltham, MA**, and is also relocating **Astra Arcus** research facility from Rochester to Boston area...**Prolifix Ltd.** team used a small peptide to inhibit the **E2F** protein complex and induced

apoptosis in mammalian tumor cells...**Vertex Pharmaceuticals, Inc.** and **Alpha Therapeutic Corp.** ended an agreement to develop VX-366 for treatment of inherited hemoglobin disorders...**NaviCyte** received Phase I SBIR grant for up to $100,000 from NIH for development of prototype of its NaviFlow technology for high-throughput screening ...**Covance Inc.** will invest $21 million in expansion and renovation of its facility in Indianapolis, IN.

Target

merely a means to an end. The critical next step is to validate the gene and its protein product as a potential drug target. The Human Genome Project continues to produce a treasure chest of expressed sequence tags (ESTs) and a tantalizing array of complete gene sequences.

Companies are applying a variety of functional genomic strategies to link genes to specific diseases and to multigenic phenotypes. Yet the ultimate challenge for pharmaceutical companies is to sift through all the sequence and differential gene expression data to identify the best targets for drug discovery.

Spinning off technology developed at the University of North Carolina (Chapel Hill), Cytogen Corp. (Princeton, NJ) formed its wholly owned subsidiary AxCell Biosciences earlier this year. The young company is building a protein interaction database, cataloging all the interactions the modular domains of proteins can engage in with a range of ligands, in order to gain insight into protein function and to select the most critical interaction to target for drug development.

AxCell's cloning-of-ligand-targets (COLT) technology employs "recognition units" from the company's genetic diversity library (GDL) to map functional protein interactions and quantitate their affinity. The company's inter-functional protcomic database (IFP-dbase) elucidates protein interaction networks and structure-activity relationships based on ligand affinity with protein modular domains.

Defining Disease Pathways

Signal Pharmaceuticals, Inc.'s (San Diego, CA) integrated drug target and discovery effort is based on mapping gene-regulating pathways in cells and identifying small molecules that regulate the activation of those genes. In collaboration with academic researchers, the company has identified a large number of regulatory proteins in several mitogen-activated protein (MAP) kinase pathways (including the JNK, FRK and p38

signaling pathways), which Signal is evaluating for the treatment of autoimmune, inflammatory, cardiovascular and neurologic diseases, and cancer. Other target identification

programs focus on the NF-kB pathway, estrogen-related genes and central/peripheral nervous system genes.

Regulating cytokine production in immune and inflammatory disorders,

The Genome Reporter Matrix depicts a subset of a yeast array. Each colony harbors a GFP-reporter construct for a single gene. Collectively, the array reports the expression of all yeast genes. A: Array in visible light. B: Image of fluorescent emission from the array.

Acacia Biosciences

and modifying bone metabolism to treat osteoporosis are the focus of Signal's collaboration with Tanabe Seiyaku (Osaka, Japan). Signal has partnered with Organon/Akzo Nobel (Netherlands) to identify estrogen-responsive genes as targets for treating neurodegenerative and psychiatric diseases, atherosclerosis and ischemia, and with Roche Bioscience (Palo Alto, CA) to develop human peripheral nerve cell lines for the discovery of treatments for pain and incontinence.

Exelixis' (S. San Francisco, CA) strategy for target selection is to define disease pathways and identify regulatory molecules that activate or inhibit those biochemical/genetic pathways. Based on the finding that these pathways are conserved across species, the company is studying the model genetic systems of Drosophila and Caenorhabditis elegans. Using its PathFinder technology, Exelixis systematically introduces mutations into the genomes of these model organisms, looking for mutations that enhance or suppress the target disease-related gene. These novel genes then become the basis of drug screening assays.

Cadus Pharmaceutical Corp. (Tarrytown, NY) is identifying surrogate ligands to newly discovered orphan G-protein coupled transmembrane receptors of unknown function to determine the suitability of the receptors as drug targets. Inserting the novel receptor in a yeast system yields a ligand that activates the receptor. Access to a surrogate ligand allows the company to screen for receptor antagonists in the yeast system.

"The antagonist plus the surrogate ligand gives you two probes—an on probe and an off probe—which allows you to look at function," explains David Webb, Ph.D., vp of research and chief scientific officer. A surrogate ligand also provides information on which G-protein interacts with the orphan receptor and its associated signaling pathways, further clarifying the role of the receptor as a potential drug target. Cadus' collaboration with SmithKline (Philadelphia) capitalizes on Cadus' ability to determine orphan receptor function, applying the technology to SmithKline's proprietary, newly discovered G-protein receptors.

Cadus' recombinant yeast system can also be used to screen cell and tissue extracts for natural ligands, and the company is accelerating its internal drug-discovery efforts in the areas of cancer, inflammation and allergy. A recent equity investment in Axiom Biotechnologies (San Diego, CA) gave Cadus a license to Axiom's high-throughput pharmacologic screening system for lead optimization and discovery.

As its name implies, gene/Networks (Alameda, CA) focuses on identifying gene networks that contribute to multigenic phenotypes and complex disease processes. The integration of mouse and human genetic studies forms the basis of the technology. The Genome Tagged Mice database in development will serve as a library of natural mouse genetic and phenotypic variation. Disease-related genes identified in mice are then evaluated in human family- and population-based studies to confirm their clinical relevance and linkages to pathophysiologic traits.

Blocking Gene Expression

Inactivating a gene known to be expressed in association with a particular disease is one approach to identifying appropriate therapeutic targets. The target validation and discovery program at Ribozyme Pharmaceuticals, Inc. (Boulder, CO) applies the company's ribozyme technology to achieve selective inhibition of gene expression in cell culture and in animals.

Correlation of the gene expression inhibition with phenotype can

---

A strong chemical combination to help you grow. And flourish.

Three hundred million dollars and ten years of hard work. That's what it costs to bring your biotechnology-derived therapeutic to the marketplace.

Which means, no room for error.

Which means, in turn, you'd be wise to tap into the combined capabilities of Mallinckrodt and J.T.Baker: dual sources, trusted names for your chemical raw materials.

Two separate GMP-produced brands offering the control of a single quality system and the convenience of a single audit process.

We offer comprehensive product lines including USP salts, bioreagents, high purity solvents and chromatography products in Beaker to Bulk™ packaging for easy scale-up.

Call 1-800-582-2537, or access our website at http://www.mallbaker.com. For dual chemical sources dedicated to helping you grow. Flourish. Succeed!

MALLINCKRODT   **J.T.Baker**

Target

suggest the relative importance of the gene in disease pathology. The company's nuclease-resistant ribozymes form the basis of a collaboration with Schering AG (Germany) for drug target validation and the development of ribozyme-based therapeutic agents, and with Chiron Corp. (Emeryville, CA) for target validation.

With several antisense compounds now progressing through clinical trials, the concept of using oligonucleotides to inhibit gene activity is not new. But rather than focusing on therapeutics development, Sequitur, Inc. (Natick, MA) is creating antisense compounds for the purpose of determining gene function and validating drug targets. Clients typically provide the one-year-old company with the sequence (or EST) of a potential gene target and, in return, Sequitur custom designs a series of three to six antisense compounds that yield a three-to-ten-fold inhibition of the target gene in cell culture. The company also provides oligofectins, a series of cationic lipids, to deliver the oligonucleotides to a variety of cultured cells.

"Differential expression information is just for correlation, it doesn't tell function or confirm what would be a good target," says Tod Woolf, Ph.D., director of technology development at Sequitur. Whereas, antisense compounds will inhibit a target, Sequitur offers both phosphorothioate DNA antisense compounds, and its proprietary Next Generation chimeric oligonucleotides, which have a higher hybridization affinity, greater specificity and reduced toxicity, according to the company.

Mining Pathogen Genomes

Companies such as Human Genome Sciences (HGS; Rockville, MD), Incyte (Palo Alto, CA),

AxCell Biosciences scientists say their technology enables the rapid and simple functional identification of the two essential molecular components of protein interaction networks: specific recognition units that bind distinct modular protein domains are identified and isolated using a combination structural/functional approach that uses both peptide phase display Generic Diversity Libraries (GDL) and bioinformatics, and cloning of Ligand Targets (COLT) technology utilizes recognition units as functional probes to isolate families of interactor proteins.

Millennium Pharmaceuticals Inc. (Cambridge, MA) and Genome Therapeutics (Waltham, MA) are relying on high-speed DNA sequencing, positional cloning and other strategies to identify specific microbial genomic sites that would be good targets for infectious disease therapeutics.

HGS recently completed sequencing of the bacterial pathogen *Streptococcus pneumoniae*, which is the focus of an agreement with Hoffmann-La Roche (Basel, Switzerland). Roche will use the sequence data to develop new anti-infectives against *S. pneumoniae*. HGS and Roche have expanded their collaboration to include a nonexclusive license to access sequence information for the intestinal bacterium *Enterococcus faecalis*.

Incyte Pharmaceuticals has completed one-fold coverage of the *Candida albicans* genome, identify-

ing 60% of the genes of this fungal pathogen. This genome will become part of the company's PathoSeq microbial database. Incyte recently introduced the ZooSeq animal gene sequence and expression database. The database will provide genomic information across various species commonly used in preclinical drug testing, which may help to better define potential drug targets.

Millennium Pharmaceuticals continues to report success in identifying novel drug targets, having recently discovered a novel chemokine called neurotactin and a new class of MAD-related proteins that inhibit transforming growth factor beta (TGF-ß) signaling. The company also received U.S. patent coverage for the tub genes, believed to play a role in obesity, and for the gene that encodes the protein melastatin, which appears to suppress metastasis in malignant melanoma. •
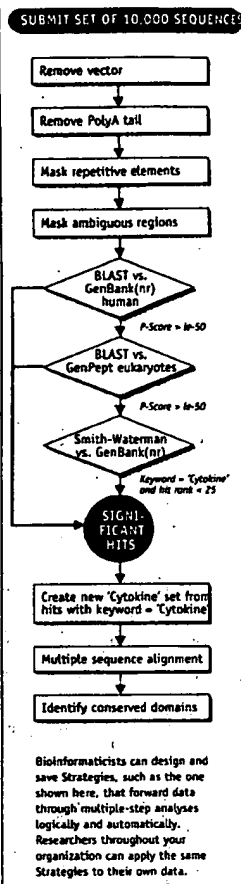

Pangea

Smith, now a computer programmer, is an expert in systems integration, Internet technologies and the application of industrial engineering principles to the drug discovery process. Before co-founding Pangea, he was the manager of software development at Attorney's Briefcase, a legal research software company.

By being "in the trenches" with customers and collaborators, Bellenson and Smith sensed the frustration of pharmaceutical researchers whose incompatible tools have impeded their progress. According to Bellenson, "Most of them are geared toward analyzing one molecule at a time. It's like emptying the ocean with an eye dropper—an incompatible eye dropper at that. A pharmaceutical company may have 30 different drug discovery teams with various approaches. The problem is to manage the process of experimenting with a lot of different approaches, to automate while maintaining flexibility."

GeneWorld 2.1 enables "integration of the entire target discovery and validation process," Bellenson says. The commercial software package coordinates the entire process of sequence-data analysis and can be integrated with other programs and databases, according to Smith, who adds that it handles thousands of sequence results, organizes and automates annotation and seamlessly interacts with growing genome databases. Simple forms and menus enable users to turn raw sequence data into crucial knowledge for drug discovery by applying algorithms to sequences, creating custom analysis strategies and producing useful reports, without the need for writing computer code. GeneWorld 2.1 runs on a variety of platforms and operating systems.

Pairing industrial relational database-management systems with a web-browser interface, Pangea's Operating System of Drug Discovery is an open-computing framework that allows client/server and Java-enabled web-based technologies to collect, organize and analyze drug discovery information for pharmaceutical companies to simplify and accelerate drug discovery. The technology unites automated genomics database analysis for drug target site selection, chemical information database analysis and large-scale combinatorial chemistry project management and high-throughput screening project management for drug lead efficacy analysis. Pangea officials maintain that these integrated elements provide a unified environment for chemists, biologists and others involved in the drug discovery process to work together with

commercial and public domain software.

Pangea's Operating System of Drug Discovery can accommodate Sybase, Oracle or Informix relational database-management systems and any version of UNIX. It absorbs new data formats, databases, algorithms and analysis paradigms into the automated workflow without software modifications. Netscape Navigator provides a friendly user interface from PC, Macintosh, and UNIX workstations.

In the near term, Pangea plans to complete its bioinformatics core with two more programs. Gene Foundry, a sample tracking and workflow sequence package for DNA sequence and fragment information, will also offer interaction with robots, reagent tracking and troubleshooting. Gene Thesaurus, the other package is a "warehouse of bioinformatics data," says Bellenson. ■

Europe

GTAC Chairman, Professor Norman C. Nevin, said 1996 saw "four important developments": an increase in enquiries and submissions made to GTAC; an increase in the complexity of submitted protocols; a continuing shift from gene therapy for single-gene disorders toward strategies aimed at tumour destruction in cancer; and a growth in international sponsorship of U.K. gene therapy trials.

Since 1993, GTAC and its predecessor, the Clothier Committee, have approved 18 U.K. gene therapy clinical trials (13 of which have been carried out), which are listed in the report. The disease areas targeted by these trials include severe combined immunodeficiency (1 trial), cystic fibrosis (6), metastatic melanoma (2), lymphoma (2), neuroblastoma (1), breast cancer (1), Hunter's syndrome (1), cervical cancer (1), glioblastoma

breast cancer, breast cancer with liver metastases, glioblastoma, malignant ascites due to gastrointestinal cancer and ovarian cancer.

Copies of the GTAC third annual report are available from the GTAC Secretariat, Wellington House, 133-155 Waterloo Road, London SE1 8UG, U.K.

Coated Lenses Prevent PCO

Scientists in the U.K. say it may be possible to prevent posterior capsule opacification (PCO), a common complication following cataract surgery, by using the implanted poly-methylmethacrylate (PMMA) intraocular lens as a drug delivery system. PCO occurs in 30-50% of cataract surgery patients as a result of stimulated cell growth within the remaining capsular bag. The condition causes a decline in visual acuity and requires expensive laser treatment, thus negating the routine use of cataract surgery in underdeveloped countries, explains G. Duncan, at the

---

Fischer-Vize, *Science* 270, 1828 (1995).
35. T. C. James and S. C. Elgin, *Mol. Cell Biol.* 6, 3862 (1986); R. Paro and D. S. Hogness, *Proc. Natl. Acad. Sci. U.S.A.* 88, 263 (1991); B. Tschiersch *et al.*, *EMBO J.* 13, 3822 (1994); M. T. Madireddi *et al.*, *Cell* 87, 75 (1996); D. G. Stokes, K. D. Tartof, R. P. Perry, *Proc. Natl. Acad. Sci. U.S.A.* 93, 7137 (1996).
36. P. M. Palosaari *et al.*, *J. Biol. Chem.* 266, 10750 (1991); A. Schmitz, K. H. Gartemann, J. Fiedler, E.

Grund, R. Eichenlaub, *Appl. Environ. Microbiol.* 58, 4068 (1992); V. Sharma, K. Suvama, R. Meganathan, M. E. Hudspeth, *J. Bacteriol.* 174, 5057 (1992); M. Kanazawa *et al.*, *Enzyme Protein* 47, 9 (1993); Z. L. Boynton, G. N. Bennet, F. B. Rudolph, *J. Bacteriol.* 178, 3015 (1996).
37. M. Ho *et al.*, *Cell* 77, 869 (1994).
38. W. Hendriks *et al.*, *J. Cell Biochem.* 59, 418 (1995).
39. We thank H. Skaletsky and F. Lewitter for help with

sequence analysis; Lawrence Livermore National Laboratory for the flow-sorted Y cosmid library; and P. Bain, A. Bortvin, A. de la Chapelle, G. Fink, K. Jegalian, T. Kawaguchi, E. Lander, H. Lodish, P. Matsudaira, D. Menke, U. RajBhandary, R. Reijo, S. Rozen, A. Schwartz, C. Sun, and C. Tilford for comments on the manuscript. Supported by NIH.

28 April 1997; accepted 9 September 1997

Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale

Joseph L. DeRisi, Vishwanath R. Iyer, Patrick O. Brown*

DNA microarrays containing virtually every gene of *Saccharomyces cerevisiae* were used to carry out a comprehensive investigation of the temporal program of gene expression accompanying the metabolic shift from fermentation to respiration. The expression profiles observed for genes with known metabolic functions pointed to features of the metabolic reprogramming that occur during the diauxic shift, and the expression patterns of many previously uncharacterized genes provided clues to their possible functions. The same DNA microarrays were also used to identify genes whose expression was affected by deletion of the transcriptional co-repressor *TUP1* or overexpression of the transcriptional activator *YAP1*. These results demonstrate the feasibility and utility of this approach to genomewide exploration of gene expression patterns.

The complete sequences of nearly a dozen microbial genomes are known, and in the next several years we expect to know the complete genome sequences of several metazoans, including the human genome. Defining the role of each gene in these genomes will be a formidable task, and understanding how the genome functions as a whole in the complex natural history of a living organism presents an even greater challenge.

Knowing when and where a gene is expressed often provides a strong clue as to its biological role. Conversely, the pattern of genes expressed in a cell can provide detailed information about its state. Although regulation of protein abundance in a cell is by no means accomplished solely by regulation of mRNA, virtually all differences in cell type or state are correlated with changes in the mRNA levels of many genes. This is fortuitous because the only specific reagent required to measure the abundance of the mRNA for a specific gene is a cDNA sequence. DNA microarrays, consisting of thousands of individual gene sequences printed in a high-density array on a glass microscope slide (*1, 2*), provide a practical and economical tool for studying gene expression on a very large scale (*3–6*).

Saccharomyces cerevisiae is an especially

Department of Biochemistry, Stanford University School of Medicine, Howard Hughes Medical Institute, Stanford, CA 94305-5428, USA.

*To whom correspondence should be addressed. E-mail: pbrown@cmgm.stanford.edu

favorable organism in which to conduct a systematic investigation of gene expression. The genes are easy to recognize in the genome sequence, cis regulatory elements are generally compact and close to the transcription units, much is already known about its genetic regulatory mechanisms, and a powerful set of tools is available for its analysis.

A recurring cycle in the natural history of yeast involves a shift from anaerobic (fermentation) to aerobic (respiration) metabolism. Inoculation of yeast into a medium rich in sugar is followed by rapid growth fueled by fermentation, with the production of ethanol. When the fermentable sugar is exhausted, the yeast cells turn to ethanol as a carbon source for aerobic growth. This switch from anaerobic growth to aerobic respiration upon depletion of glucose, referred to as the diauxic shift, is correlated with widespread changes in the expression of genes involved in fundamental cellular processes such as carbon metabolism, protein synthesis, and carbohydrate storage (*7*). We used DNA microarrays to characterize the changes in gene expression that take place during this process for nearly the entire genome, and to investigate the genetic circuitry that regulates and executes this program.

Yeast open reading frames (ORFs) were amplified by the polymerase chain reaction (PCR), with a commercially available set of primer pairs (*8*). DNA microarrays, containing approximately 6400 distinct DNA sequences, were printed onto glass slides by

using a simple robotic printing device (*9*). Cells from an exponentially growing culture of yeast were inoculated into fresh medium and grown at 30°C for 21 hours. After an initial 9 hours of growth, samples were harvested at seven successive 2-hour intervals, and mRNA was isolated (*10*). Fluorescently labeled cDNA was prepared by reverse transcription in the presence of Cy3(green)- or Cy5(red)-labeled deoxyuridine triphosphate (dUTP) (*11*) and then hybridized to the microarrays (*12*). To maximize the reliability with which changes in expression levels could be discerned, we labeled cDNA prepared from cells at each successive time point with Cy5, then mixed it with a Cy3-labeled "reference" cDNA sample prepared from cells harvested at the first interval after inoculation. In this experimental design, the relative fluorescence intensity measured for the Cy3 and Cy5 fluors at each array element provides a reliable measure of the relative abundance of the corresponding mRNA in the two cell populations (Fig. 1). Data from the series of seven samples (Fig. 2), consisting of more than 43,000 expression-ratio measurements, were organized into a database to facilitate efficient exploration and analysis of the results. This database is publicly available on the Internet (*13*).

During exponential growth in glucose-rich medium, the global pattern of gene expression was remarkably stable. Indeed, when gene expression patterns between the first two cell samples (harvested at a 2-hour interval) were compared, mRNA levels differed by a factor of 2 or more for only 19 genes (0.3%), and the largest of these differences was only 2.7-fold (*14*). However, as glucose was progressively depleted from the growth media during the course of the experiment, a marked change was seen in the global pattern of gene expression. mRNA levels for approximately 710 genes were induced by a factor of at least 2, and the mRNA levels for approximately 1030 genes declined by a factor of at least 2. Messenger RNA levels for 183 genes increased by a factor of at least 4, and mRNA levels for 203 genes diminished by a factor of at least 4. About half of these differentially expressed genes have no currently recognized function and are not yet named. Indeed, more than 400 of the differentially expressed genes have no apparent homology

to any gene whose function is known (15). The responses of these previously uncharacterized genes to the diauxic shift therefore provides the first small clue to their possible roles.

The global view of changes in expression of genes with known functions provides a vivid picture of the way in which the cell adapts to a changing environment. Figure 3 shows a portion of the yeast metabolic pathways involved in carbon and energy metabolism. Mapping the changes we observed in the mRNAs encoding each enzyme onto this framework allowed us to infer the redirection in the flow of metabolites through this system. We observed large inductions of the genes coding for the enzymes aldehyde dehydrogenase (ALD2) and acetyl–coenzyme A(CoA) synthase (ACS1), which function together to convert the products of alcohol dehydrogenase into acetyl-CoA, which in turn is used to fuel the tricarboxylic acid (TCA) cycle and the glyoxylate cycle. The concomitant shutdown of transcription of the genes encoding pyruvate decarboxylase and induction of pyruvate carboxylase rechannels pyruvate away from acetaldehyde, and instead to oxalacetate, where it can serve to supply the TCA cycle and gluconeogenesis. Induction of the pivotal genes PCK1, encoding phosphoenolpyruvate carboxykinase, and FBP1, encoding fructose 1,6-biphosphatase, switches the directions of two key irreversible steps in glycolysis, reversing the flow of metabolites along the reversible steps of the glycolytic pathway toward the essential biosynthetic precursor, glucose-6-phosphate. Induction of the genes coding for the trehalose synthase and glycogen synthase complexes promotes channeling of glucose-6-phosphate into these carbohydrate storage pathways.

Just as the changes in expression of genes encoding pivotal enzymes can provide insight into metabolic reprogramming, the behavior of large groups of functionally related genes can provide a broad view of the systematic way in which the yeast cell adapts to a changing environment (Fig. 4). Several classes of genes, such as cytochrome c–related genes and those involved in the TCA/glyoxylate cycle and carbohydrate storage, were coordinately induced by glucose exhaustion. In contrast, genes devoted to protein synthesis, including ribosomal proteins, tRNA synthetases, and translation, elongation, and initiation factors, exhibited a coordinated decrease in expression. More than 95% of ribosomal genes showed at least twofold decreases in expression during the diauxic shift (Fig. 4) (13). A noteworthy and illuminating exception was that the

genes encoding mitochondrial ribosomal genes were generally induced rather than repressed after glucose limitation, highlighting the requirement for mitchondrial biogenesis (13). As more is learned about the functions of every gene in the yeast genome, the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns will become increasingly powerful.

Several distinct temporal patterns of expression could be recognized, and sets of genes could be grouped on the basis of the similarities in their expression patterns. The characterized members of each of these groups also shared important similarities in their functions. Moreover, in most cases, common regulatory mechanisms could be inferred for sets of genes with similar expression profiles. For example, seven genes showed a late induction profile, with mRNA levels increasing by more than ninefold at

the last timepoint but less than threefold at the preceding timepoint (Fig. 5B). All of these genes were known to be glucose-repressed, and five of the seven were previously noted to share a common upstream activating sequence (UAS), the carbon source response element (CSRE) (16–20). A search in the promoter regions of the remaining two genes, ACR1 and IDP2, revealed that ACR1, a gene essential for ACS1 activity, also possessed a consensus CSRE motif, but interestingly, IDP2 did not. A search of the entire yeast genome sequence for the consensus CSRE motif revealed only four additional candidate genes, none of which showed a similar induction.

Examples from additional groups of genes that shared expression profiles are illustrated in Fig. 5, C through F. The sequences upstream of the named genes in Fig. 5C all contain stress response elements (STRE), and with the exception
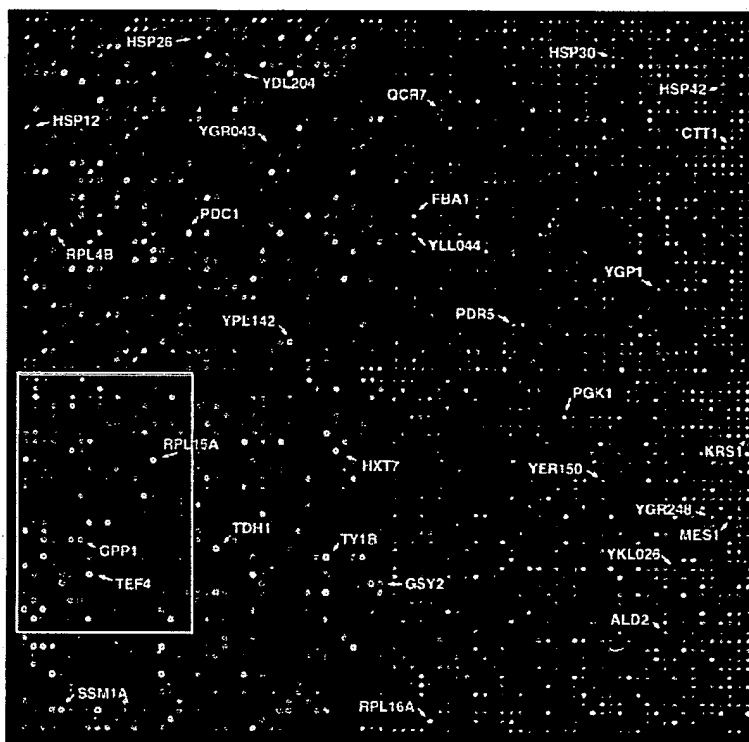

Fig. 1. Yeast genome microarray. The actual size of the microarray is 18 mm by 18 mm. The microarray was printed as described (9). This image was obtained with the same fluorescent scanning confocal microscope used to collect all the data we report (49). A fluorescently labeled cDNA probe was prepared from mRNA isolated from cells harvested shortly after inoculation (culture density of <5 × 10⁶ cells/ml and media glucose level of 19 g/liter) by reverse transcription in the presence of Cy3-dUTP. Similarly, a second probe was prepared from mRNA isolated from cells taken from the same culture 9.5 hours later (culture density of ~2 × 10⁸ cells/ml, with a glucose level of <0.2 g/liter) by reverse transcription in the presence of Cy5-dUTP. In this image, hybridization of the Cy3-dUTP–labeled cDNA (that is, mRNA expression at the initial timepoint) is represented as a green signal, and hybridization of Cy5-dUTP–labeled cDNA (that is, mRNA expression at 9.5 hours) is represented as a red signal. Thus, genes induced or repressed after the diauxic shift appear in this image as red and green spots, respectively. Genes expressed at roughly equal levels before and after the diauxic shift appear in this image as yellow spots.

of HSP42, have previously been shown to be controlled at least in part by these elements (21–24). Inspection of the sequences upstream of HSP42 and the two uncharacterized genes shown in Fig. 5C, YKL026c, a hypothetical protein with similarity to glutathione peroxidase, and YGR043c, a putative transaldolase, revealed that each of these genes also possess repeated upstream copies of the stress-responsive CCCCT motif. Of the 13 additional genes in the yeast genome that shared this expression profile [including HSP30, ALD2, OM45, and 10 uncharacterized ORFs (25)], nine contained one or more recognizable STRE sites in their upstream regions.

The heterotrimeric transcriptional activator complex HAP2,3,4 has been shown to be responsible for induction of several genes important for respiration (26–28). This complex binds a degenerate consensus sequence known as the CCAAT box (26). Computer analysis, using the consensus sequence TNRYTGGB (29), has suggested that a large number of genes involved in respiration may be specific targets of HAP2,3,4 (30). Indeed, a putative HAP2,3,4 binding site could be found in the sequences upstream of each of the seven cytochrome c–related genes that showed the greatest magnitude of induction (Fig. 5D). Of 12 additional cytochrome c–related genes that were induced, HAP2,3,4 binding sites were present in all but one. Significantly, we found that transcription of HAP4 itself was induced nearly ninefold concomitant with the diauxic shift.

Control of ribosomal protein biogenesis is mainly exerted at the transcriptional level, through the presence of a common upstream-activating element (UAS_{rpg}) that is recognized by the Rap1 DNA-binding protein (31, 32). The expression profiles of seven ribosomal proteins are shown in Fig. 5F. A search of the sequences upstream of all seven genes revealed consensus Rap1-binding motifs (33). It has been suggested that declining Rap1 levels in the cell during starvation may be responsible for the decline in ribosomal protein gene expression (34). Indeed, we observed that the abundance of RAP1 mRNA diminished by 4.4-fold, at about the time of glucose exhaustion.

Of the 149 genes that encode known or putative transcription factors, only two, HAP4 and SIP4, were induced by a factor of more than threefold at the diauxic shift. SIP4 encodes a DNA-binding transcriptional activator that has been shown to interact with Snf1, the "master regulator" of glucose repression (35). The eightfold induction of SIP4 upon depletion of glucose strongly suggests a role in the induction of
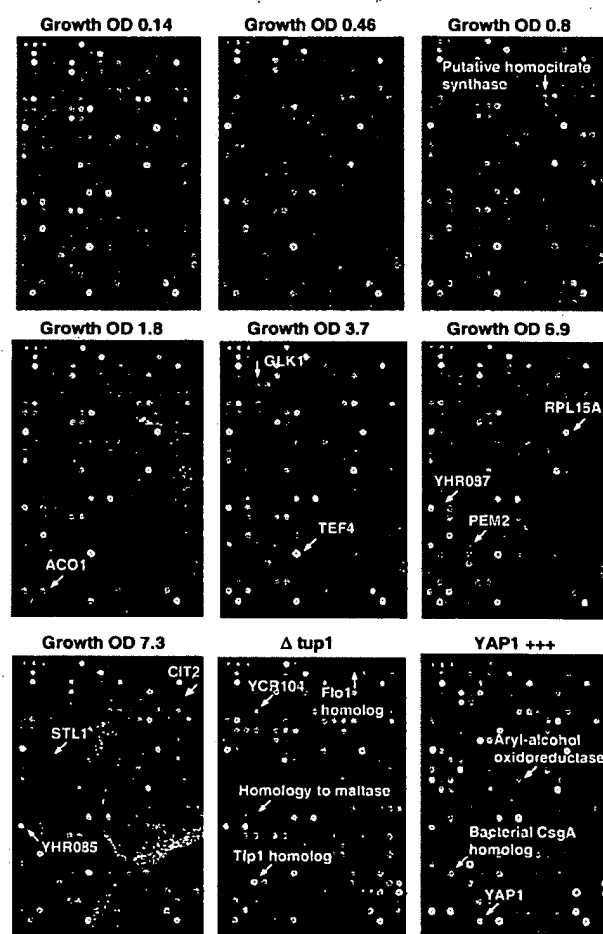
downstream genes at the diauxic shift.

Although most of the transcriptional responses that we observed were not previously known, the responses of many genes during the diauxic shift have been described. Comparison of the results we obtained by DNA microarray hybridization with previously reported results therefore provided a strong test of the sensitivity and accuracy of this approach. The expression patterns we observed for previously characterized genes showed almost perfect concordance with previously published results (36). Moreover, the differential expression measurements obtained by DNA microarray hybridization were reproducible in duplicate experiments. For example, the remarkable changes in gene expression between cells harvested immediately after inoculation and immediately after the diauxic shift (the first and sixth intervals in this time series) were measured in duplicate, independent DNA microarray hybridizations. The correlation coefficient for two complete sets of expression ratio measurements was 0.87, and for more than 95% of the genes, the expres-

sion ratios measured in these duplicate experiments differed by less than a factor of 2. However, in a few cases, there were discrepancies between our results and previous results, pointing to technical limitations that will need to be addressed as DNA microarray technology advances (37, 38). Despite the noted exceptions, the high concordance between the results we obtained in these experiments and those of previous studies provides confidence in the reliability and thoroughness of the survey.

The changes in gene expression during this diauxic shift are complex and involve integration of many kinds of information about the nutritional and metabolic state of the cell. The large number of genes whose expression is altered and the diversity of temporal expression profiles observed in this experiment highlight the challenge of understanding the underlying regulatory mechanisms. One approach to defining the contributions of individual regulatory genes to a complex program of this kind is to use DNA microarrays to identify genes whose expression is affected

Fig. 2. The section of the array indicated by the gray box in Fig. 1 is shown for each of the experiments described here. Representative genes are labeled. In each of the arrays used to analyze gene expression during the diauxic shift, red spots represent genes that were induced relative to the initial timepoint, and green spots represent genes that were repressed relative to the initial timepoint. In the arrays used to analyze the effects of the tup1Δ mutation and YAP1 overexpression, red spots represent genes whose expression was increased, and green spots represent genes whose expression was decreased by the genetic modification. Note that distinct sets of genes are induced and repressed in the different experiments. The complete images of each of these arrays can be viewed on the Internet (13). Cell density as measured by optical density (OD) at 600 nm was used to measure the growth of the culture.

by mutations in each putative regulatory gene. As a test of this strategy, we analyzed the genomewide changes in gene expression that result from deletion of the *TUP1* gene. Transcriptional repression of many genes by glucose requires the DNA-binding repressor

Mig1 and is mediated by recruiting the transcriptional co-repressors Tup1 and Cyc8/Ssn6 (*39*). Tup1 has also been implicated in repression of oxygen-regulated, mating-type–specific, and DNA-damage–inducible genes (*40*).

Wild-type yeast cells and cells bearing a deletion of the *TUP1* gene (*tup1*Δ) were grown in parallel cultures in rich medium containing glucose as the carbon source. Messenger RNA was isolated from exponentially growing cells from the two populations and used to prepare cDNA labeled with Cy3 (green) and Cy5 (red), respectively (*11*). The labeled probes were mixed and simultaneously hybridized to the microarray. Red spots on the microarray therefore represented genes whose transcription was induced in the *tup1*Δ strain, and thus presumably repressed by Tup1 (*41*). A representative section of the microarray (Fig. 2, bottom middle panel) illustrates that the genes whose expression was affected by the *tup1*Δ mutation, were, in general, distinct from those induced upon glucose exhaustion [complete images of all the arrays shown in Fig. 2 are available on the Internet (*13*)]. Nevertheless, 34 (10%) of the genes that were induced by a factor of at least 2 after the diauxic shift were similarly induced by deletion of *TUP1*, suggesting that these genes may be subject to *TUP1*-mediated repression by glucose. For example, *SUC2*, the gene encoding invertase, and all five hexose transporter genes that were induced during the course of the diauxic shift were similarly induced, in duplicate experiments, by the deletion of *TUP1*.

The set of genes affected by Tup1 in this experiment also included α-glucosidases, the mating-type–specific genes *MFA1* and *MFA2*, and the DNA damage–inducible *RNR2* and *RNR4*, as well as genes involved in flocculation and many genes of unknown function. The hybridization signal corresponding to expression of *TUP1* itself was also severely reduced because of the (incomplete) deletion of the transcription unit in the *tup1*Δ strain, providing a positive control in the experiment (*42*).

Many of the transcriptional targets of Tup1 fell into sets of genes with related biochemical functions. For instance, although only about 3% of all yeast genes appeared to be *TUP1*-repressed by a factor of more than 2 in duplicate experiments under these conditions, 6 of the 13 genes that have been implicated in flocculation (*15*) showed a reproducible increase in expression of at least twofold when *TUP1* was deleted. Another group of related genes that appeared to be subject to *TUP1* repression encodes the serine-rich cell wall mannoproteins, such as Tip1 and Tir1/Srp1 which are induced by cold shock and other stresses (*43*), and similar, serine-poor proteins, the seripauperins (*44*). Messenger RNA levels for 23 of the 26 genes in this group were reproducibly elevated by at least 2.5-fold in the *tup1*Δ
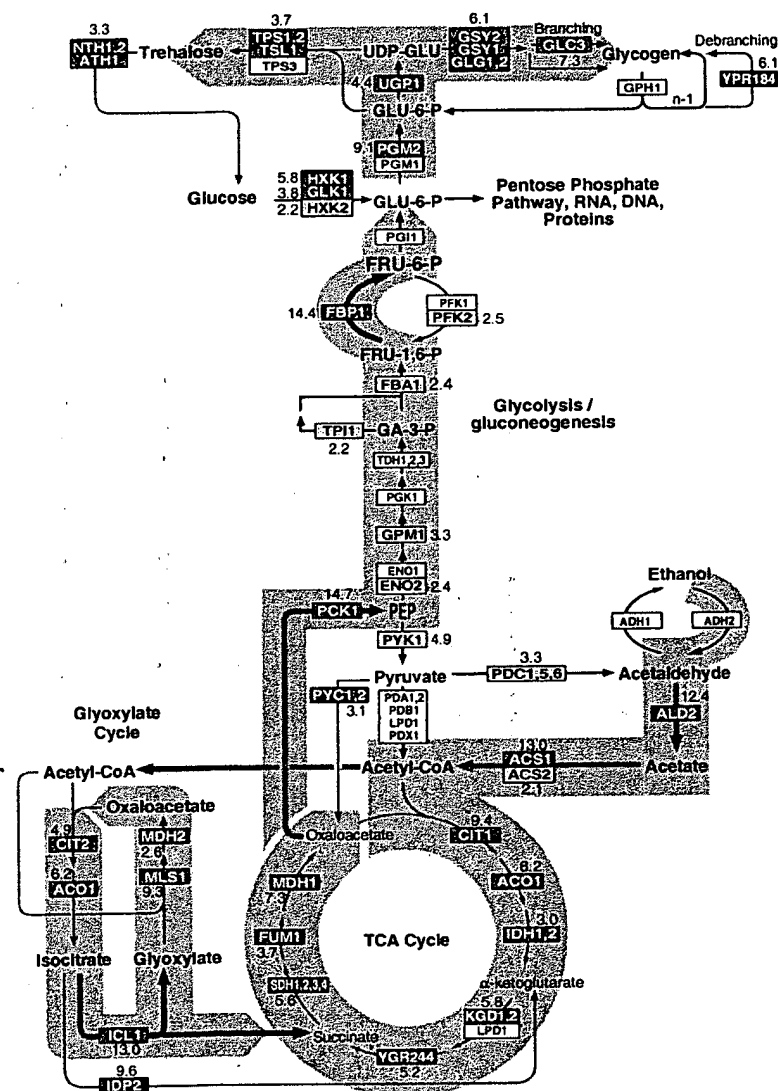

Fig. 3. Metabolic reprogramming inferred from global analysis of changes in gene expression. Only key metabolic intermediates are identified. The yeast genes encoding the enzymes that catalyze each step in this metabolic circuit are identified by name in the boxes. The genes encoding succinyl-CoA synthase and glycogen-debranching enzyme have not been explicitly identified, but the ORFs YGR244 and YPR184 show significant homology to known succinyl-CoA synthase and glycogen-debranching enzymes, respectively, and are therefore included in the corresponding steps in this figure. Red boxes with white lettering identify genes whose expression increases in the diauxic shift. Green boxes with dark green lettering identify genes whose expression diminishes in the diauxic shift. The magnitude of induction or repression is indicated for these genes. For multimeric enzyme complexes, such as succinate dehydrogenase, the indicated fold-induction represents an unweighted average of all the genes listed in the box. Black and white boxes indicate no significant differential expression (less than twofold). The direction of the arrows connecting reversible enzymatic steps indicate the direction of the flow of metabolic intermediates, inferred from the gene expression pattern, after the diauxic shift. Arrows representing steps catalyzed by genes whose expression was strongly induced are highlighted in red. The broad gray arrows represent major increases in the flow of metabolites after the diauxic shift, inferred from the indicated changes in gene expression.

strain, and 18 of these genes were induced by more than sevenfold when *TUP1* was deleted. In contrast, none of 83 genes that could be classified as putative regulators of the cell division cycle were induced more than twofold by deletion of *TUP1*. Thus, despite the diversity of the regulatory systems that employ Tup1, most of the genes that it regulates under these conditions fall into a limited number of distinct functional classes.

Because the microarray allows us to monitor expression of nearly every gene in yeast, we can, in principle, use this approach to identify all the transcriptional targets of a regulatory protein like Tup1. It is important to note, however, that in any single experiment of this kind we can only recognize those target genes that are normally repressed (or induced) under the conditions of the experiment. For instance, the experiment described here analyzed a *MAT* α strain in which *MFA1* and *MFA2*, the genes encoding the a-factor mating pheromone precursor, are normally repressed. In the isogenic *tup1*Δ strain, these genes were inappropriately expressed, reflecting the role that Tup1 plays in their repression. Had we instead carried out this experiment with a *MATA* strain (in which expression of *MFA1* and *MFA2* is not repressed), it would not have been possible to conclude anything regarding the role of Tup1 in the repression of these genes. Conversely, we cannot distinguish indirect effects of the chronic absence of Tup1 in the mutant strain from effects directly attributable to its participation in repressing the transcription of a gene.

Another simple route to modulating the activity of a regulatory factor is to overexpress the gene that encodes it. *YAP1* encodes a DNA-binding transcription factor belonging to the b-zip class of DNA-binding proteins. Overexpression of *YAP1* in yeast confers increased resistance to hydrogen peroxide, o-phenanthroline, heavy metals, and osmotic stress (45). We analyzed differential gene expression between a wild-type strain bearing a control plasmid and a strain with a plasmid expressing *YAP1* under the control of the strong *GAL1-10* promoter, both grown in galactose (that is, a condition that induces *YAP1* overexpression). Complementary DNA from the control and *YAP1* overexpressing strains, labeled with Cy3 and Cy5, respectively, was prepared from mRNA isolated from the two strains and hybridized to the microarray. Thus, red spots on the array represent genes that were induced in the strain overexpressing *YAP1*.

Of the 17 genes whose mRNA levels increased by more than threefold when

YAP1 was overexpressed in this way, five bear homology to aryl-alcohol oxidoreductases (Fig. 2 and Table 1). An additional four of the genes in this set also belong to the general class of dehydrogenases/oxidoreductases. Very little is known about the role of aryl-alcohol oxidoreductases in *S. cerevisiae*, but these enzymes have been isolated from ligninolytic fungi, in which they participate in coupled redox reactions, oxidizing aromatic, and aliphatic unsaturated alcohols to aldehydes with the production of hydrogen peroxide (46, 47). The fact that a remarkable fraction of the targets identified in this experiment belong to the same small, functional group of oxidoreductases suggests that these genes

might play an important protective role during oxidative stress. Transcription of a small number of genes was reduced in the strain overexpressing Yap1. Interestingly, many of these genes encode sugar permeases or enzymes involved in inositol metabolism.

We searched for Yap1-binding sites (TTACTAA or TGACTAA) in the sequences upstream of the target genes we identified (48). About two-thirds of the genes that were induced by more than threefold upon Yap1 overexpression had one or more binding sites within 600 bases upstream of the start codon (Table 1), suggesting that they are directly regulated by Yap1. The absence of canonical Yap1-bind-

Fig. 4. Coordinated regulation of functionally related genes. The curves represent the average induction or repression ratios for all the genes in each indicated group. The total number of genes in each group was as follows: ribosomal proteins, 112; translation elongation and initiation factors, 25; tRNA synthetases (excluding mitochondial synthetases), 17; glycogen and trehalose synthesis and degradation, 15; cytochrome c oxidase and reductase proteins, 19; and TCA- and glyoxylate-cycle enzymes, 24.
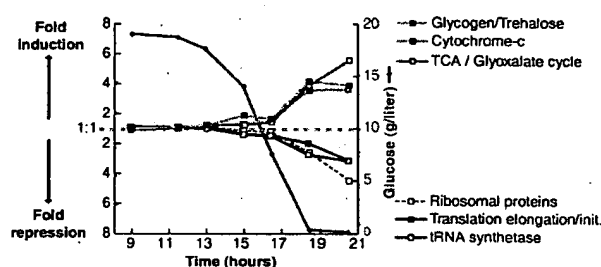

Table 1. Genes induced by *YAP1* overexpression. This list includes all the genes for which mRNA levels increased by more than twofold upon *YAP1* overexpression in both of two duplicate experiments, and for which the average increase in mRNA level in the two experiments was greater than threefold (50). Positions of the canonical Yap1 binding sites upstream of the start codon, when present, and the average fold-increase in mRNA levels measured in the two experiments are indicated.

ORF	Distance of Yap1 site from ATG	Gene	Description	Fold-increase
YNL331C			Putative aryl-alcohol reductase	12.9
YKL071W	162–222 (5 sites)		Similarity to bacterial csgA protein	10.4
YML007W	.	YAP1	Transcriptional activator involved in oxidative stress response	9.8
YFL056C	223, 242		Homology to aryl-alcohol dehydrogenases	9.0
YLL060C	98		Putative glutathione transferase	7.4
YOL165C	266		Putative aryl-alcohol dehydrogenase (NADP+)	7.0
YCR107W			Putative aryl-alcohol reductase	6.5
YML116W	409	ATR1	Aminotriazole and 4-nitroquinoline resistance protein	6.5
YBR008C	142, 167, 364		Homology to benomyl/methotrexate resistance protein	6.1
YCLX08C			Hypothetical protein	6.1
YJR155W			Putative aryl-alcohol dehydrogenase	6.0
YPL171C	148, 212	OYE3	NAPDH dehydrogenase (old yellow enzyme), isoform 3	5.8
YLR460C	167, 317		Homology to hypothetical proteins YCR102c and YNL134c	4.7
YKR076W	178		Homology to hypothetical protein YMR251w	4.5
YHR179W	327	OYE2	NAD(P)H oxidoreductase (old yellow enzyme), isoform 1	4.1
YML131W	507		Similarity to *A. thaliana* zeta-crystallin homolog	3.7
YOL126C		MDH2	Malate dehydrogenase	3.3

ing sites upstream of the others may reflect an ability of Yap1 to bind sites that differ from the canonical binding sites, perhaps in cooperation with other factors, or less likely, may represent an indirect effect of Yap1 overexpression, mediated by one or more intermediary factors. Yap1 sites were found only four times in the corresponding region of an arbitrary set of 30 genes that were not differentially regulated by Yap1.

Use of a DNA microarray to characterize the transcriptional consequences of mutations affecting the activity of regulatory molecules provides a simple and powerful approach to dissection and characterization of regulatory pathways and net-

works. This strategy also has an important practical application in drug screening. Mutations in specific genes encoding candidate drug targets can serve as surrogates for the ideal chemical inhibitor or modulator of their activity. DNA microarrays can be used to define the resulting signature pattern of alterations in gene expression, and then subsequently used in an assay to screen for compounds that reproduce the desired signature pattern.

DNA microarrays provide a simple and economical way to explore gene expression patterns on a genomic scale. The hurdles to extending this approach to any other organism are minor. The equipment

required for fabricating and using DNA microarrays (9) consists of components that were chosen for their modest cost and simplicity. It was feasible for a small group to accomplish the amplification of more than 6000 genes in about 4 months and, once the amplified gene sequences were in hand, only 2 days were required to print a set of 110 microarrays of 6400 elements each. Probe preparation, hybridization, and fluorescent imaging are also simple procedures. Even conceptually simple experiments, as we described here, can yield vast amounts of information. The value of the information from each experiment of this kind will progressively increase as more is learned about the functions of each gene and as additional experiments define the global changes in gene expression in diverse other natural processes and genetic perturbations. Perhaps the greatest challenge now is to develop efficient methods for organizing, distributing, interpreting, and extracting insights from the large volumes of data these experiments will provide.

REFERENCES AND NOTES

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Science 270, 467 (1995).
2. D. Shalon, S. J. Smith, P. O. Brown, Genome Res. 6, 639 (1996).
3. D. Lashkari, Proc. Natl. Acad. Sci. U.S.A., in press.
4. J. DeRisi et al., Nature Genet. 14, 457 (1996).
5. D. J. Lockhart et al., Nature Biotechnol. 14, 1675 (1996).
6. M. Chee et al., Science 274, 610 (1996).
7. M. Johnston and M. Carlson, in The Molecular Biology of the Yeast Saccharomyces: Gene Expression, E. W. Jones, J. R. Pringle, J. R. Broach, Eds. (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1992), p. 193.
8. Primers for each known or predicted protein coding sequence were supplied by Research Genetics. PCR was performed with the protocol supplied by Research Genetics, using genomic DNA from yeast strain S288C as a template. Each PCR product was verified by agarose gel electrophoresis and was deemed correct if the lane contained a single band of appropriate mobility. Failures were marked as such in the database. The overall success rate for a single-pass amplification of 6116 ORFs was ~94.5%.
9. Glass slides (Gold Seal) were cleaned for 2 hours in a solution of 2 N NaOH and 70% ethanol. After rinsing in distilled water, the slides were then treated with a 1:5 dilution of poly-L-lysine adhesive solution (Sigma) for 1 hour, and then dried for 5 min at 40°C in a vacuum oven. DNA samples from 100-µl PCR reactions were purified by ethanol purification in 96-well microtiter plates. The resulting precipitates were resuspended in 3× standard saline citrate (SSC) and transferred to new plates for arraying. A custom-built arraying robot was used to print on a batch of 110 slides. Details of the design of the microarrayer are available at cmgm.stanford.edu/pbrown. After printing, the microarrays were rehydrated for 30 s in a humid chamber and then snap-dried for 2 s on a hot plate (100°C). The DNA was then ultraviolet (UV)-crosslinked to the surface by subjecting the slides to 60 mJ of energy (Stratagene Stratalinker). The rest of the poly-L-lysine surface was blocked by a 15-min incubation in a solution of 70 mM succinic anhydride dissolved in a solution consisting of 315 ml of 1-methyl-2-pyrrolidinone (Aldrich) and 35 ml of 1 M boric acid (pH 8.0). Directly after the blocking reac-
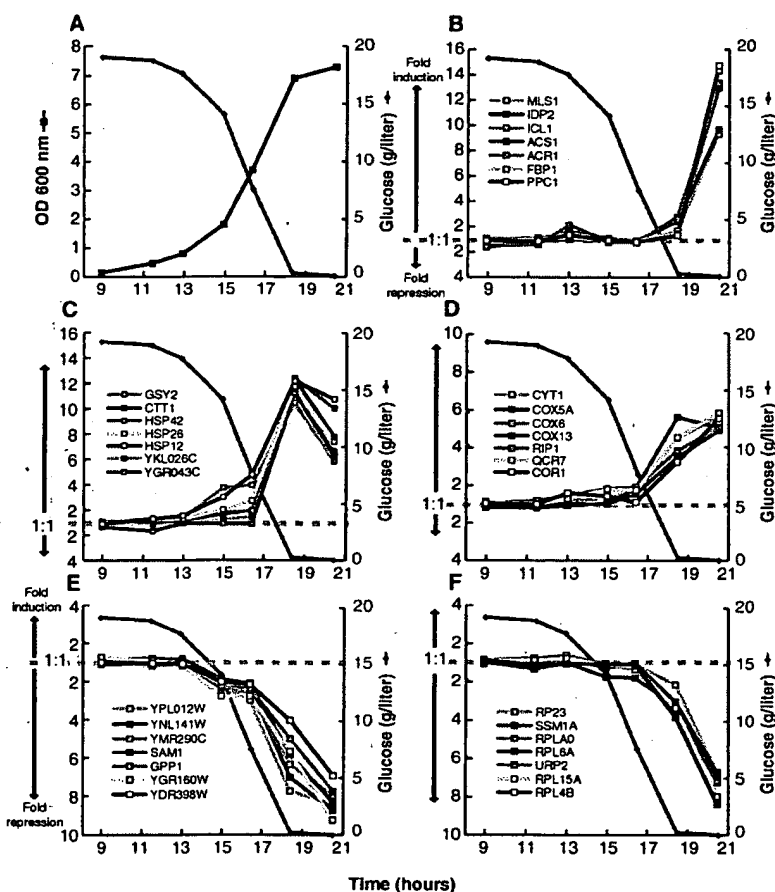
Fig. 5. Distinct temporal patterns of induction or repression help to group genes that share regulatory properties. (A) Temporal profile of the cell density, as measured by OD at 600 nm and glucose concentration in the media. (B) Seven genes exhibited a strong induction (greater than ninefold) only at the last timepoint (20.5 hours). With the exception of IDP2, each of these genes has a CSRE UAS. There were no additional genes observed to match this profile. (C) Seven members of a class of genes marked by early induction with a peak in mRNA levels at 18.5 hours. Each of these genes contain STRE motif repeats in their upstream promoter regions. (D) Cytochrome c oxidase and ubiquinol cytochrome c reductase genes. Marked by an induction coincident with the diauxic shift, each of these genes contains a consensus binding motif for the HAP2,3,4 protein complex. At least 17 genes shared a similar expression profile. (E) SAM1, GPP1, and several genes of unknown function are repressed before the diauxic shift, and continue to be repressed upon entry into stationary phase. (F) Ribosomal protein genes comprise a large class of genes that are repressed upon depletion of glucose. Each of the genes profiled here contains one or more RAP1-binding motifs upstream of its promoter. RAP1 is a transcriptional regulator of most ribosomal proteins.

tion, the bound DNA was denatured by a 2-min incubation in distilled water at ~95°C. The slides were then transferred into a bath of 100% ethanol at room temperature, rinsed, and then spun dry in a clinical centrifuge. Slides were stored in a closed box at room temperature until used.

10. YPD medium (8 liters), in a 10-liter fermentation vessel, was inoculated with 2 ml of a fresh overnight culture of yeast strain DBY7286 (MATa, ura3, GAL2). The fermentor was maintained at 30°C with constant agitation and aeration. The glucose content of the media was measured with a UV test kit (Boehringer Mannheim, catalog number 716251) Cell density was measured by OD at 600-nm wavelength. Aliquots of culture were rapidly withdrawn from the fermentation vessel by peristaltic pump, spun down at room temperature, and then flash frozen with liquid nitrogen. Frozen cells were stored at −80°C.

11. Cy3-dUTP or Cy5-dUTP (Amersham) was incorporated during reverse transcription of 1.25 μg of polyadenylated [poly(A)+] RNA, primed by a dT(16) oligomer. This mixture was heated to 70°C for 10 min, and then transferred to ice. A premixed solution, consisting of 200 U Superscript II (Gibco), buffer, deoxyribonucleoside triphosphates, and fluorescent nucleotides, was added to the RNA. Nucleotides were used at these final concentrations: 500 μM for dATP, dCTP, and dGTP and 200 μM for dTTP. Cy3-dUTP and Cy5-dUTP were used at a final concentration of 100 μM. The reaction was then incubated at 42°C for 2 hours. Unincorporated fluorescent nucleotides were removed by first diluting the reaction mixture with of 470 μl of 10 mM tris-HCl (pH 8.0)/1 mM EDTA and then subsequently concentrating the mix to ~5 μl, using Centricon-30 microconcentrators (Amicon).

12. Purified, labeled cDNA was resuspended in 11 μl of 3.5× SSC containing 10 μg poly(dA) and 0.3 μl of 10% SDS. Before hybridization, the solution was boiled for 2 min and then allowed to cool to room temperature. The solution was applied to the microarray under a cover slip, and the slide was placed in a custom hybridization chamber which was subsequently incubated for ~8 to 12 hours in a water bath at 62°C. Before scanning, slides were washed in 2× SSC, 0.2% SDS for 5 min, and then 0.05× SSC for 1 min. Slides were dried before scanning by centrifugation at 500 rpm in a Beckman CS-6R centrifuge.

13. The complete data set is available on the Internet at cmgm.stanford.edu/pbrown/explore/index.html

14. For 95% of all the genes analyzed, the mRNA levels measured in cells harvested at the first and second interval after inoculation differed by a factor of less than 1.5. The correlation coefficient for the comparison between mRNA levels measured for each gene in these two different mRNA samples was 0.98. When duplicate mRNA preparations from the same cell sample were compared in the same way, the correlation coefficient between the expression levels measured for the two samples by comparative hybridization was 0.99.

15. The numbers and identities of known and putative genes, and their homologies to other genes, were gathered from the following public databases: Saccharomyces Genome Database (genome-www. stanford.edu), Yeast Protein Database (quest7. proteome.com), and Munich Information Centre for Protein Sequences (speedy.mips.biochem.mpg.de/ mips/yeast/index.htmlx).

16. A. Scholer and H. J. Schuller, Mol. Cell. Biol. 14, 3613 (1994).

17. S. Kratzer and H. J. Schuller, Gene 161, 75 (1995).

18. R. J. Haselbeck and H. L. McAlister, J. Biol. Chem. 268, 12116 (1993).

19. M. Fernandez, E. Fernandez, R. Rodicio, Mol. Gen. Genet. 242, 727 (1994).

20. A. Hartig et al., Nucleic Acids Res. 20, 5677 (1992).

21. P. M. Martinez et al., EMBO J. 15, 2227 (1996).

22. J. C. Varela, U. M. Praekelt, P. A. Meacock, R. J. Planta, W. H. Mager, Mol. Cell. Biol. 15, 6232 (1995).

23. H. Ruis and C. Schuller, Bioessays 17, 959 (1995).

24. J. L. Parrou, M. A. Teste, J. Francois, Microbiology 143, 1891 (1997).

25. This expression profile was defined as having an induction of greater than 10-fold at 18.5 hours and less than 11-fold at 20.5 hours.

26. S. L. Forsburg and L. Guarente, Genes Dev. 3, 1166 (1989).

27. J. T. Olesen and L. Guarente, ibid. 4, 1714 (1990).

28. M. Rosenkrantz, C. S. Kell, E. A. Pennell, L. J. Devenish, Mol. Microbiol. 13, 119 (1994).

29. Single-letter abbreviations for the amino acid residues are as follows: A, Ala; C, Cys; D, Asp; E, Glu; F, Phe; G, Gly; H, His; I, Ile; K, Lys; L, Leu; M, Met; N, Asn; P, Pro; Q, Gln; R, Arg; S, Ser; T, Thr; V, Val; W, Trp; and Y, Tyr. The nucleotide codes are as follows: B—C, G, or T; N—G, A, T, or C; R—A or G; and Y—C or T.

30. C. Fondrat and A. Kalogeropoulos, Comput. Appl. Biosci. 12, 363 (1996).

31. D. Shore, Trends Genet. 10, 408 (1994).

32. R. J. Planta and H. A. Raue, ibid. 4, 64 (1988).

33. The degenerate consensus sequence VYCYRNNC-MNH was used to search for potential RAP1-binding sites. The exact consensus, as defined by (30), is WACAYCCRTACATYW, with up to three differences allowed.

34. S. F. Neuman, S. Bhattacharya, J. R. Broach, Mol. Cell. Biol. 15, 3187 (1995).

35. P. Lesage, X. Yang, M. Carlson, ibid. 16, 1921 (1996).

36. For example, we observed large inductions of the genes coding for PCK1, FBP1 [Z. Yin et al., Mol. Microbiol. 20, 751 (1996)], the central glyoxylate cycle gene ICL1 [A. Scholer and H. J. Schuller, Curr. Genet. 23, 375 (1993)], and the "aerobic" isoform of acetyl-CoA synthase, ACS1 [M. A. van den Berg et al., J. Biol. Chem. 271, 28953 (1996)], with concomitant down-regulation of the glycolytic-specific genes PYK1 and PFK2 [P. A. Moore et al., Mol. Cell. Biol. 11, 5330 (1991)]. Other genes not directly involved in carbon metabolism but known to be induced upon nutrient limitation include genes encoding cytosolic catalase T CTT1 [P. H. Bissinger et al., ibid. 9, 1309 (1989)] and several genes encoding small heat-shock proteins, such as HSP12, HSP26, and HSP42 [I. Farkas et al., J. Biol. Chem. 266, 15602 (1991); U. M. Praekelt and P. A. Meacock, Mol. Gen. Genet. 223, 97 (1990); D. Wotton et al., J. Biol. Chem. 271, 2717 (1996)].

37. The levels of induction we measured for genes that were expressed at very low levels in the uninduced state (notably, FBP1 and PCK1) were generally lower than those previously reported. This discrepancy was likely due to the conservative background subtraction method we used, which generally resulted in overestimation of very low expression levels (46).

38. Cross-hybridization of highly related sequences can also occasionally obscure changes in gene expression, an important concern where members of gene families are functionally specialized and differentially regulated. The major alcohol dehydrogenase genes, ADH1 and ADH2, share 88% nucleotide identity. Reciprocal regulation of these genes is an important feature of the diauxic shift, but was not observed in this experiment, presumably because of cross-hybridization of the fluorescent cDNAs representing these two genes. Nevertheless, we were able to detect differential expression of closely related isoforms of other enzymes, such as HXK1/HXK2 (77% identical) [P. Herrero et al., Yeast 11, 137 (1995)], MLS1/ DAL7 (73% identical) (20), and PGM1/PGM2 (72% identical) [D. Oh, J. E. Hopper, Mol. Cell. Biol. 10, 1415 (1990)], in accord with previous studies. Use in the microarray of deliberately selected DNA sequences corresponding to the most divergent segments of homologous genes, in lieu of the complete gene sequences, should relieve this problem in many cases.

39. F. E. Williams, U. Varanasi, R. J. Trumbly, Mol. Cell. Biol. 11, 3307 (1991).

40. D. Tzamarias and K. Struhl, Nature 369, 758 (1994).

41. Differences in mRNA levels between the tup1Δ and wild-type strain were measured in two independent experiments. The correlation coefficient between the complete sets of expression ratios measured in these duplicate experiments was 0.83. The concordance between the sets of genes that appeared to be induced was very high between the two experiments. When only the 355 genes that showed at least a twofold increase in mRNA in the tup1Δ strain in either of the duplicate experiments were compared, the correlation coefficient was 0.82.

42. The tup1Δ mutation consists of an insertion of the LEU2 coding sequence, including a stop codon, between the ATG of TUP1 and an Eco R I site 124 base pairs before the stop codon of the TUP1 gene.

43. L. R. Kowalski, K. Kondo, M. Inouye, Mol. Microbiol. 15, 341 (1995).

44. M. Viswanathan, G. Muthukumar, Y. S. Cong, J. Lenard, Gene 148, 149 (1994).

45. D. Hirata, K. Yano, T. Miyakawa, Mol. Gen. Genet. 242, 250 (1994).

46. A. Gutierrez, L. Caramelo, A. Prieto, M. J. Martinez, A. T. Martinez, Appl. Environ. Microbiol. 60, 1783 (1994).

47. A. Muheim et al., Eur. J. Biochem. 195, 369 (1991).

48. J. A. Wemmie, M. S. Szczypka, D. J. Thiele, W. S. Moye-Rowley, J. Biol. Chem. 269, 32592 (1994).

49. Microarrays were scanned using a custom-built scanning laser microscope built by S. Smith with software written by N. Ziv. Details concerning scanner design and construction are available at cmgm. stanford.edu/pbrown. Images were scanned at a resolution of 20 μm per pixel. A separate scan, using the appropriate excitation line, was done for each of the two fluorophores used. During the scanning process, the ratio between the signals in the two channels was calculated for several array elements containing total genomic DNA. To normalize the two channels with respect to overall intensity, we then adjusted photomultiplier and laser power settings such that the signal ratio at these elements was as close to 1.0 as possible. The combined images were analyzed with custom-written software. A bounding box, fitted to the size of the DNA spots in each quadrant, was placed over each array element. The average fluorescent intensity was calculated by summing the intensities of each pixel present in a bounding box, and then dividing by the total number of pixels. Local area background was calculated for each array element by determining the average fluorescent intensity for the lower 20% of pixel intensities. Although this method tends to underestimate the background, causing an underestimation of extreme ratios, it produces a very consistent and noise-tolerant approximation. Although the analog-to-digital board used for data collection possesses a wide dynamic range (12 bits), several signals were saturated (greater than the maximum signal intensity allowed) at the chosen settings. Therefore, extreme ratios at bright elements are generally underestimated. A signal was deemed significant if the average intensity after background subtraction was at least 2.5-fold higher than the standard deviation in the background measurements for all elements on the array.

50. In addition to the 17 genes shown in Table 1, three additional genes were induced by an average of more than threefold in the duplicate experiments, but in one of the two experiments, the induction was less than twofold (range 1.6- to 1.9-fold).

51. We thank H. Bennett, P. Spellman, J. Ravetto, M. Eisen, R. Pillai, B. Dunn, T. Ferea, and other members of the Brown lab for their assistance and helpful advice. We also thank S. Friend, D. Botstein, S. Smith, J. Hudson, and D. Dolginow for advice, support, and encouragement; K. Struhl and S. Chatterjee for the Tup1 deletion strain; L. Fernandes for helpful advice on Yap1; and S. Klapholz and the reviewers for many helpful comments on the manuscript. Supported by a grant from the National Human Genome Research Institute (NHGRI) (HG00450), and by the Howard Hughes Medical Institute (HHMI). J.D.R. was supported by the HHMI and the NHGRI. V.R. was supported in part by an Institutional Training Grant in Genome Science (T32 HG00044) from the NHGRI. P.O.B. is an associate investigator of the HHMI.

5 September 1997; accepted 22 September 1997